



American Society of
Agricultural and Biological Engineers

Go Back

Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria

D. N. Moriasi, M. W. Gitau, N. Pai, P. Daggupati

Published in *Transactions of the ASABE* 58(6): 1763-1785 (doi: 10.13031/trans.58.10715).
2015 American Society of Agricultural and Biological Engineers.

Submitted for review in April 2014 as manuscript number NRES 10715; approved for publication by the Natural Resources & Environmental Systems Community of ASABE in December 2015.

Mention of company or trade names is for description only and does not imply endorsement by the USDA. The USDA is an equal opportunity provider and employer.

The authors are **Daniel N. Moriasi, ASABE Member**, Hydrologist, USDA-ARS Grazinglands Research Laboratory, El Reno, Oklahoma; **Margaret W. Gitau, ASABE Member**, Associate Professor, Department of Agricultural and Biological Engineering, Purdue University, West Lafayette, Indiana; **Naresh Pai, ASABE Member**, Environmental Modeler, Stone Environmental, Inc., Montpelier, Vermont; **Prasad Daggupati, ASABE Member**, Assistant Research Scientist, Texas Agrilife Research, Texas A&M University College Station, Texas. **Corresponding author:** Daniel N. Moriasi, USDA-ARS Grazinglands Research Laboratory, 7207 W. Cheyenne St., El Reno, OK 73036-0000; phone: 405-262-5291; e-mail: daniel.moriasi@ars.usda.gov.

Abstract. ♦ Performance measures (PMs) and corresponding performance evaluation criteria (PEC) are important aspects of calibrating and validating hydrologic and water quality models and should be updated with advances in modeling science. We synthesized PMs and PEC from a previous special collection, performed a meta-analysis of performance data reported in recent peer-reviewed literature for three widely published watershed-scale models (SWAT, HSPF, WARMF), and one field-scale model (ADAPT), and provided guidelines for model performance evaluation²²² = ♦² = ♦² = ♦² = ♦s and more data become available.

Keywords. Guidelines, Model calibration and validation, Performance measures and evaluation criteria.

Hydrologic and water quality (H/WQ) models are increasingly being used to determine the impacts of land management, land use, climate, and conservation practices on water resources, ecology, and water-related

ecosystem services. Hydrologic cycle components and fate and transport of sediments and chemicals are examples of complex systems comprised of many processes that can be simulated using H/WQ models. A majority of H/WQ models require some degree of calibration to reduce the uncertainty of predictions (Engel et al., 2007; USEPA, 2009). Calibration is the process of adjusting input parameter values and initial or boundary conditions within reasonable ranges until the simulated results closely match the observed variables (Zeckoski et al., 2015). Calibration requires the examination of accuracy of outputs and process simulation (Sorooshian, 1983) to ensure adequate watershed and scenario representation. This requires use of model performance measures (PMs) and the corresponding performance evaluation criteria (PEC). Throughout this article, the term \diamond PMs \diamond refers to the statistical and graphical methods used during model calibration and validation, \diamond performance data \diamond refers to the reported values of each of the statistical PMs (e.g., 0.5 for NSE), and \diamond PEC \diamond refers to model performance qualitative ratings (e.g., very good, good, satisfactory, or unsatisfactory) with the corresponding quantitative thresholds for the statistical PMs of interest (e.g., NSE, PBIAS, or R^2). Validation is the process by which a calibrated model is shown to be capable of reproducing a set of field observations or predicting future conditions without further adjustment to the calibrated parameters (Zheng et al., 2012).

Modelers have used different PMs, including statistical, graphical, or a combination of both. For example, Herr and Chen (2012) preferred the use of absolute and relative error, while Huth et al. (2012) recommended and used a variety of measures, including Nash-Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970) and the ratio of root mean square error (RMSE) and standard deviation of measured data (RSR; Moriasi et al., 2007). Commonly used graphical PMs include time series plots (e.g., van der Keur et al., 2001; Mutiti and Levy, 2010; Palosuo et al., 2011; Arnold et al., 2012; Herr and Chen, 2012; Huth et al., 2012), scatter plots (e.g., Palosuo et al., 2011; Herr and Chen, 2012), cumulative charts (e.g., Herr and Chen, 2012), and contour maps (e.g., Zheng et al., 2012).

Nevertheless, the use of both graphical and statistical PMs is essential for robust model performance evaluation (Biondi et al., 2012; Bennett et al., 2013; Harmel et al., 2014; Daggupati et al., 2015a). For instance, measures such as the NSE are insensitive to systematic errors and yield good model performance even if low values are poorly fitted (Krause et al., 2005; Ritter and Muñoz-Carpena, 2013; Pfannerstill et al., 2014). In such cases, graphical PMs provide supplementary evidence as to where (e.g., in the time series, magnitude of event, depth, etc.) the model is not performing adequately. In addition, pre-inspection of graphical output likely minimizes equifinality (or parameter non-uniqueness), a situation in which a variety of parameter sets can yield acceptable model performance (Beven and Freer, 2001; Doherty and Johnston, 2003). This is achieved by allowing identification of parameter sets that provide better process simulation, thereby reducing the number of possible parameter sets that yield acceptable model performance. Recent works indicate that the intended use of the model could serve as an important factor in the selection of PMs and PEC (Finsterle et al., 2012; Harmel et al., 2014).

Past literature on model PMs includes Willmott (1984), Loague and Green (1991), ASCE (1993), Refsgaard (1997), Gupta et al. (1998), Legates and McCabe (1999), Santhi et al. (2001), Krause (2005), McCuen et al. (2006), Engel et al. (2007), and Moriasi et al. (2007). With respect to PEC, several studies have provided a summary of ranges of values for use in assessing model performance (Popov, 1979; Ramanarayanan et al., 1997; Gassman et al., 2007; Moriasi et al., 2007; Douglas-Mankin et al., 2010; Tuppad et al., 2011; Ritter and Muñoz-Carpena, 2013). The use of PEC provides objective indications of the adequacy of model performance, hence affording greater credibility to the modeling work (Duda et al., 2012). General PEC help model users and decision makers estimate model calibration and validation accuracy, usability for their specific application, and uncertainty or reliability of model predictions (Duda et al., 2012). It is also important to set PEC before beginning model evaluation (ASCE, 1993; USEPA, 2002; Engel, 2007; Moriasi et al., 2007).

Selection and use of PEC also varies by study and by model (Santhi et al., 2001; Van Liew et al., 2007; Parajuli et al., 2009; Bennett et al., 2013; Daggupati et al., 2014; Harmel et al., 2014). This could result in inconsistent model evaluation, making it difficult to provide a benchmark for further model improvements.

Moriasi et al. (2007) provided guidance to facilitate a more consistent and structured approach for model performance evaluation. However, the scope of the guidelines provided by Moriasi et al. (2007) was limited to NSE, percent bias (PBIAS; Gupta et al., 1999), and RSR for stream flow, sediment, and nutrient (N and P) simulations at a monthly temporal scale and watershed spatial scale. Different PMs can have differing ranges of conditions for which they are best suited (Krause et al., 2005; Gupta et al., 2009; Westerberg et al., 2011; Pushpalatha et al., 2012). Just as there are differences in PMs, there are also differences in the PEC for each measure. In addition, models perform differently for different simulated response outputs and, perhaps, at different temporal and spatial scales (Westerberg et al., 2011; Biondi et al., 2012), which may require different PEC. For example, regions with a shallow water table (e.g., south Florida) experience rapid water table rise within 12 hours of rainfall or irrigation input (Jaber et al., 2006; Hendricks et al., 2013). Hendricks et al. (2013) evaluated a daily temporal scale model for simulating water table responses in a shallow water table region of Florida and concluded that a daily temporal scale was a fundamental limitation because the hydrologic response time was less than 12 hours. Therefore, there is need to explore how different models perform under different conditions using different PMs to help determine appropriate PEC. Further, Moriasi et al. (2007) stated that “as new and improved methods and information are developed, the recommended guidelines should be updated to reflect these developments.”

Recently, Biondi et al. (2012), Ritter and Muñoz-Carpena (2013), Moriasi et al. (2012), Pushpalatha et al. (2012), Bennett et al. (2013), Black et al. (2014), and Harmel et al. (2014) focused on various aspects of performance of H/WQ models. Biondi et al. (2012) performed a literature review and provided general model validation guidelines that cover several topics discussed in this special collection. Black et al. (2014) provided general guidance on the implementation and application of water resource management models focused on scenario analysis. Bennett et al. (2013) reviewed and provided methods available across different fields for describing the performance of environmental models focusing on model PMs. Pushpalatha et al. (2012) analyzed several forms of NSE to determine the form that was suitable for flows. Ritter and Muñoz-Carpena (2013) presented a unified framework for determining model PEC in a statistically rigorous way and for the evaluation of bias, outliers, and repeated data focused on RMSE and NSE. Harmel et al. (2014) reviewed literature and recommended a broad methodology that takes into account intended use to establish model performance expectations. The methodology provides a brief summary of several topics, including model valuation, interpretation, and communication of model results.

Moriasi et al. (2012) summarized the results of 25 H/WQ models in a special collection of 22 articles, each focusing on individual model calibration and validation strategies. The special collection provided a good source of model-specific calibration and validation examples, performance evaluation examples, and references. However, there is need for consistent model calibration and validation guidelines (Moriasi et al., 2012), including PMs and PEC.

Recognizing the good work done by others, in this article we: (1) synthesize the special collection articles (Moriasi et al., 2012) with respect to PMs and PEC; (2) perform a meta-analysis of performance data as reported in peer-reviewed literature by considering the effects of calibration and validation periods, simulated components, and spatial and temporal scales; and (3) establish guidelines for model performance evaluation based on information from the synthesis (objective 1) and meta-analysis (objective 2). Further, we present an example case study illustrating the application of our recommendations in model calibration and validation.

In summary, this article is one of nine topic-specific articles in a special collection whose main goal is to provide recommendations, which together with information from other literature will be used to develop model calibration and validation engineering practices for H/WQ models. These articles extensively cover critical issues related to the calibration and validation of H/WQ models. This article focuses on model PMs and the corresponding PEC related to models in the Moriasi et al. (2012) special collection and provides a more rigorous framework than Moriasi et al. (2007, 2012) for determining PEC, involving a meta-analysis of the performance data collected in this study and using the results to guide PEC development.

Methods

Synthesis of Performance Measures and Evaluation Criteria

As a starting point, the articles in the Moriassi et al. (2012) special collection were reviewed to determine the statistical and graphical PMs used for each of the models. The models in the special collection were grouped into three spatial categories (point to plot, field, and watershed; table 1). PMs and PEC reported outside of the special collection were helpful in broadening the outlook on PEC and providing additional materials useful for establishing guidelines. Commonly used PMs and PEC within and outside the special collection (Moriassi et al., 2012) for each model were recorded for in-depth analyses.

Table 1. Models in the Moriassi et al. (2012) special collection grouped by spatial scale.			
Model		Simulated Processes (Components)	Reference
Point to plot scale			
	COUPMODEL	Hydrology, N, carbon, plant growth, heat, tracer, chloride	Jansson (2012)
	HYDRUS	Water flow, solute transport, heat transfer, carbon dioxide	Jimunek et al. (2012)
	MACRO	Macropore flow, pesticides	Jarvis and Larsbo (2012)
	MT3DMS	Multispecies solute transport, groundwater	Zheng et al. (2012)
	SHAW	Hydrology, heat transfer	Flerchinger et al. (2012)
	STANMOD	Solute transport in soils and groundwater	van Genuchten et al. (2012)
	SWIM3	Water and solute movement	Huth et al. (2012)
	TOUGH2	Multiphase, multicomponent fluids in porous and fractured geologic media	Finsterle et al. (2012)
	VS2DI	Water, solute, heat transport	Healy and Essaid (2012)
Field scale			
	ADAPT	Hydrology, erosion, nutrients, pesticides, subsurface tile drainage	Gowda et al. (2012)
	CREAMS/GLEAMS	Hydrology, erosion, pesticides, sediments, nutrients, plant growth	Knisel and Douglas-Mankin (2012)
	DAISY	Water, snowmelt, carbon cycle, energy balance, N cycle, crop production, pesticides	Hansen et al. (2012)

	DRAINMOD	Hydrology (water table depth, tile flow, surface runoff, depth of irrigation water applied, wetland hydrology), plant growth (crop yield)	Skaggs et al. (2012)
	EPIC/APEX	Hydrology (surface runoff, stream flow, tile flow), plant growth, erosion, sediments, nutrients, pesticides	Wang et al. (2012)
	RZWQM2	Hydrology, plant growth, nutrients, pesticides	Ma et al. (2012)
	WEPP Hillslope	Hydrology, soil erosion	Flanagan et al. (2012)
Watershed scale			
	BASINS/HSPF	Hydrology, snowmelt, pollutant loadings, erosion, fate and transport	Duda et al. (2012)
	KINEROS2/AGWA	Runoff, erosion, sediments	Goodrich et al. (2012)
	MIKE-SHE	Surface and subsurface water dynamics, interception, evapotranspiration, overland flow, channel flow, unsaturated flow, saturated zone flow, water levels, surface and groundwater quality	Jaber and Shukla (2012)
	SWAT	Hydrology, plant growth, sediments, nutrients, pesticides	Arnold et al. (2012)
	WAM	Hydrology, sediments, nutrients	Bottcher et al. (2012)
	WARMF	Hydrology, sediments, nutrients, acid mine, carbon, bacteria	Herr and Chen (2012)
	WEPP Watershed	Hydrology, soil erosion	Flanagan et al. (2012)

Although there are several ways in which statistical PMs can be categorized (Moriassi et al., 2007; Bennett et al., 2013), in this article statistical PMs are discussed and divided into three broad categories: (1) standard regression, (2) dimensionless, and (3) error index based on Moriassi et al. (2007). Standard regression measures determine the strength of the linear relationship between simulated and measured data.

Dimensionless measures provide a relative model evaluation assessment, and error index measures quantify the deviation in the units of the data of interest (Legates and McCabe, 1999). Graphical PMs are divided into two categories (direct and derived comparison), and information about the strengths and weaknesses of each of the measures was obtained from the literature. In this article, we define direct comparison graphical PMs as graphical PMs in which original measured and simulated data are compared with each other, for instance, with time series graphs. Derived graphical PMs are those in which measured or simulated data are first transformed into another form before they are displayed in a comparative graph, for example, frequency duration curves.

A comparative analysis of the reported PMs was performed to evaluate (1) how they compare across the models, (2) their advantages and disadvantages, and (3) their usability (ease of and suitability for use) from a user or non-developer perspective. Additional considerations for PMs included their suitability for event-based vs. continuous models and their use with missing and/or discrete observed data. Based on this analysis, recommendations are made for suitable PMs.

Meta-Analysis of Performance Data

A statistical meta-analysis was performed on the model performance data to guide the development of the PEC. Simply stated, a meta-analysis (Glass, 1976; Hunter et al., 1982; Hunt, 1997; Lyons, 1998; among others) is the accumulation and analysis of data from separate but similar studies for the purpose of obtaining insights from the pooled data that are not discernible from the individual studies. The methodology provides an avenue for bringing together information from various related studies in search of common patterns and conclusions. It can also be used to reconcile data from disparate studies. Since its inception in the 1970s, meta-analysis has been applied successfully in various fields, including medical research and social studies (Egger and Smith, 1997; Lyons, 1998; Bland, 2000). The methodology has also been used successfully in natural resources and environmental systems for the development of a Best Management Practice (BMP) tool (Gitau et al., 2005).

The accumulation of data from existing studies is the most involved part of a meta-analysis, as it requires considerable attention to some key considerations, as described in ensuing subsections.

Kinds of Articles to Include

It is necessary that articles be relevant to the study at hand (Light and Smith, 1971; Hunt, 1997) and that the articles contain the information needed to achieve study goals. As materials may be subject to re-interpretation, it is preferable that the articles contain original material and include a detailed account of the study. Further, given a common tendency toward selecting articles that favor an author's viewpoint and/or that align with prevailing opinion (Egger and Smith, 1997), it is important that article selection follows an objective procedure. For example, in this article, the articles included are primary sources that provided performance data for the various PMs. Additional criteria included the presence of details such as models used, evaluation time step, components evaluated, and whether data reported were for calibration or validation.

Whether or Not to Use Only Published Material

Generally, published material is deemed to have more reliable data and is afforded more credibility than unpublished material. However, published material is often preferential in nature, favoring research works based on reported significance (Lipsey and Wilson, 2001). For example, in regard to model performance, articles reporting higher values of NSE may be preferentially published, whereas those with lower values (albeit with better parameter representations) may take a while longer or may not be published at all. Including only published material may result in a publication bias (Light and Smith, 1971; Hunter et al., 1982; Light and Pillemer, 1984; Bland, 2000); thus, we recommend that both published and unpublished material be included. The challenge lies in being able to find unpublished information, as this is not generally

available. Thus, the dataset developed for this article only contains data from published material (peer-reviewed journal articles after 1990).

Rejection of Articles on the Basis of Perceived Inadequacies in the Methodology

Another important consideration is the determination of article suitability for inclusion based on methodologies used. This is especially so for unpublished information, as a work may be unpublished due to unsuitable methodologies. However, it is important to note that flaws can be identified in almost any article (Hunter et al., 1982; Lipsey and Wilson, 2001) given that opinions tend to differ among researchers. The use of methodology as a basis for article inclusion would thus introduce elements of subjectivity into the analysis (Light and Smith, 1971) and would result in a reduced dataset (Glass, 1976; Hunter et al., 1982), which would then impact the analysis. In this study, no judgements were made as to the adequacy or inadequacy of the methodologies used once an article was deemed suitable for inclusion based on study goals.

Amount of Data Necessary for Analyses

The ideal case would be to have all existing data; in this case, the details and results of all studies in which model calibration and validation have been conducted and performance values have been reported. However, this is generally not practical, due to limited access to unpublished material, if nothing else, and thus the need for a representative sample arises. In addition, it is necessary to consider the study goals. For example, in this article, the goal was to capture recent advances in modeling (in the 1990s and later) for commonly used H/WQ models published in a recent special collection (Moriassi et al., 2012) when establishing performance criteria. For this work, the target was to review a minimum of 20 articles (outside the Moriassi et al. (2012) special collection) per model for the most commonly simulated output responses (flow, sediment, and nutrients) to be reviewed. To enable meta-analysis, each reported entry of performance data was extracted and tabulated along with size of the study area (supplemental material tables S1-1 through S1-22, available at http://bit.ly/NRES_SW10715). Exceptions were permitted for models for which the available peer-reviewed articles numbered less than 20, in which case all available articles were reviewed. Data on stream flow, surface runoff, base flow, and tile flow model performance values were combined as appropriate and referred to as flow for the watershed-scale and ADAPT models to ensure that there were sufficient data for analyses. Where stream flow was the only component used in the analysis and/or discussion, the term **stream flow** was used to distinguish it from the combined flow component. Data were commonly reported in the literature at annual, monthly, and daily temporal scales for watershed-scale models and at a monthly temporal scale for field-scale models. In addition, there was a substantial amount of seasonal data associated with PBIAS.

Handling of Extreme Values

Values showing up as extreme values, once all data are assembled, may reflect extreme site or study characteristics; thus, their exclusion would mask the existence of extremes. Therefore, extreme values such as values of other PMs for studies in which there were negative NSE values were not excluded from the primary analysis. However, negative NSE values were not included in criteria development, as such values represent unacceptable model performance. Further description is provided under the **Meta-Analysis of Performance Data** subheading within the **Results and Discussion** section.

Data Analyses

Once all data are assembled, the most basic analysis involves determining an average for each data component (Hunter et al., 1982; Light and Pillimer, 1984; Hunt, 1997), for example, an average of all NSE values. More detailed approaches involve the computation of standardized metrics to account for differences in the amounts of data among studies (Light and Pillimer, 1984; Lipsey and Wilson, 2001). In either case, this would mask the variability in the data, so more in-depth analysis allowing the examination of factors that could affect results (Hunter et al., 1982; Light and Pillimer, 1984; Hunt, 1997) and extraction of other

pertinent information are necessary.

In this study, descriptive statistics such as mean, median, minimum, and maximum were computed for the performance data, and the associated distributions were plotted in order to make a determination on subsequent analysis. Following these preliminary diagnostics, significant differences in reported values were determined based on (1) calibration or validation; (2) scale (specifically watershed-scale studies based on Hydrologic Unit Code (HUC; <https://pubs.er.usgs.gov/publication/ofr84708>; direct comparisons were not made between watershed and field scales due to the large difference in available data); and (3) model components (e.g., flow, sediment, and nutrients). The analysis was conducted using the median test, a non-parametric (typically distribution-free) test based on median rank scores (SAS, 2007; Sheskin, 2003; Brown and Mood, 1951). The test considers all observations and ranks them as 0 or 1 based on their location around (above or below) the median. Resulting rank scores are then used for the comparisons based on the chi-square statistic and associated probabilities. In addition, the performance data were plotted on a common axis to provide a visual comparison. All analyses were carried out using JMP statistical software (SAS, 2008).

Development of Guidelines for Model Performance Evaluation

The median test on reported performance data was used to determine whether separate PEC were needed for calibration and validation periods, spatial and temporal scales, and for different simulated response outputs. Following the median test, thresholds for model PEC ratings were established by computing percentiles or quartiles of model PM data collected from peer-reviewed articles outside the Moriasi et al. (2012) special collection. The thresholds obtained for the defined ratings formed the initial PEC, which along with the results of the synthesis of the PEC and the modeling experience of the authors were used to develop final PEC guidelines for identified separate categories. A similar approach was used by USEPA (2010) as part of an evaluation of the potential benefits of numeric nutrient criteria for Florida's flowing waters. The guidelines are in the form of recommended PMs and PEC. Brief descriptions are provided for (1) the importance of following proper calibration and validation procedures (Zeckoski et al., 2015; Arnold et al., 2015; Baffaut et al., 2015; Malone et al., 2015; Daggupati et al., 2015b; Guzman et al., 2015; and Yuan et al., 2015) prior to using these general guidelines; (2) additional considerations for adjusting the general recommendations because of the variety of modeling applications; and (3) a framework for determining recommended model PMs and their corresponding PEC.

Results and Discussion

Synthesis of Performance Measures and Evaluation Criteria

The most commonly used graphical PMs in the special collection articles were time series charts (table 2; e.g., WARMF, DAISY, VS2DI, SWIM3, and SWAT). Other graphical PMs included scatter plots (e.g., APEX/EPIC, CREAMS/GLEAMS, DAISY, WARMF, and SWAT), cumulative frequency curves (e.g., WARMF, SWAT), contour maps (e.g., MT3DMS), depth profile plots (e.g., SWIM3), thermographs in which heat is used as a surrogate for water movement (e.g., VS2DI), and bar charts (e.g., EPIC/APEX). Thermographs are quite common in soil/ water-solute transport applications.

The most commonly used statistical PMs were NSE, RMSE (also called root mean square deviation, RMSD), and R^2 (table 2). Other reported statistical PMs included d (Willmott, 1981), PBIAS (Gupta et al., 1999), mean absolute error, R, absolute error, relative error, standard error of estimate, non-parametric tests, RSR (Moriasi et al., 2007), 95% confidence intervals (to account for uncertainty, mean, and standard deviation), autocorrelation, and cross-correlation (table 2). Brief descriptions as well as discussions of the strengths, weaknesses, and usage of the commonly used measures are presented in ensuing subsections. The abbreviations of the models in the Moriasi et al. (2012) special collection are provided in the Appendix, while

the statistical PMs and associated equations are provided in table 5. Detailed accounts of these and other measures can be obtained from model-specific articles and in the literature (e.g., Wilmott, 1984; Legates and McCabe, 1999; Krause et al., 2005; Moriasi et al., 2007; Ritter and Muñoz-Carpena, 2013; Bennett et al., 2013; Harmel et al., 2014).

Of the models within the Moriasi et al. (2012) special collection, only a few provided PEC (table 3), including BASINS/HSPF (Duda et al., 2012), DRAINMOD (Skaggs et al., 2012), EPIC/APEX (Wang et al., 2012), KINEROS/AGWA (Goodrich et al., 2012), RZWQM2 (Ma et al., 2012), and WARMF (Herr and Chen, 2012). PEC from Moriasi et al. (2007) were cited for SWAT (Arnold et al., 2012), SWIM3 (Huth et al., 2012), and WEPP (Flanagan et al., 2012). With the exception of SWIM3 (Huth et al., 2012), all point and plot scale models (table 3) employed user-defined objective function thresholds with autocalibration algorithms (Moriasi et al., 2012). The MIKE-SHE (Jaber and Shukla, 2012) and WAM (Bottcher et al., 2012) articles do not provide any PEC.

Table 2. Summary of performance measures and evaluation criteria for H/WQ models in the Moriasi et al. (2012) special collection.

Model	Suggested Performance Measures and Evaluation Criteria							
	Statistical Performance Measures ^[a]						Performance Evaluation Criteria ^[b]	Graphical Performance Measures ^[c]
	NSE	R ²	RMSE	d	PBIAS	Other		
Point to plot scale								
COUPMODEL	X	X	-	-	-	-	n.p.	Time series
HYDRUS	-	X	-	-	-	X	n.p.	Time series
MACRO	X	-	X	-	-	-	n.p.	-
MT3DMS	-	-	-	-	-	X	n.p.	Contour maps, time series
SHAW	-	-	X	-	-	-	n.p.	Time series
STANMOD	-	-	-	-	-	X	n.p.	Time series
SWIM3	X	-	-	-	-	X	Moriasi et al. (2007)	Time series
TOUGH2	-	-	-	-	-	X	n.p.	-
VS2DI	-	-	-	-	-	X	n.p.	Time series
Field scale								
ADAPT	X	-	X	X	-	X	n.p.	Time series, scatter plots
CREAMS/GLEAMS	X	X	-	X	-	X	n.p.	Time series
DAISY	-	-	X	X	-	-	n.p.	Scatter plots
DRAINMOD	X	X	-	-	-	X	Table 3	Time series
EPIC/APEX	X	X	X	-	X	X	Table 3	Time series, scatter plots, bar charts
RZWQM2	-	-	X	-	-	-	Table 3	Time series

	WEPP Hillslope	X	-	X	-	X	X	Moriasi et al. (2007)	-
	Watershed scale								
	BASINS/HSPF	-	X	-	-	-	X	Table 3	Time series, scatter plots, CFC
	KINEROS2/AGWA	X	-	-	-	-	X	Table 3	Time series
	MIKE-SHE	-	-	X	X	-	-	n.p.	Time series
	SWAT	X	X	X	-	X	X	Moriasi et al. (2007)	Time series, scatter plots, CFC
	WAM	X	-	X	-	-	-	n.p.	Time series
	WARMF	-	-	-	-	-	X	Table 3	Time series, scatter plots, CFC
	WEPP Watershed	X	-	X	-	X	X	Moriasi et al. (2007)	-
<p>[a] ♦♦♦NSE = Nash Sutcliffe efficiency/coefficient, R^2 = coefficient of determination, RMSE = root mean square error/deviation, d = index of agreement, PBIAS = percent bias/deviation.</p> <p>♦Other♦ includes root mean square error to standard deviation ratio, linear or weighted correlation coefficient, mean error, mean absolute error, standard error of estimate, 95% confidence interval, comparison between observed and predicted means and standard deviations, mean and variance of weighted residuals, autocorrelation, cross-correlation, nonparametric tests, t-tests, and objective functions.</p> <p>[b] ♦♦♦n.p. = not provided and user-defined.</p> <p>[c] ♦♦♦CFC = cumulative frequency curves.</p>									

Graphical Performance Measures

Graphical PMs provide an important complementary tool for modelers to support the calibration and validation of H/WQ models (Daggupati et al., 2015a). Graphical PMs allow visual comparison of simulated and measured output response data, help identify model bias, identify differences in timing and magnitude of peaks (e.g., peak flows) and shape of recession curves, incorporate measurement (Harmel and Smith, 2007) and model (Shirmohammadi et al., 2006) uncertainty, and illustrate how well the model reproduces the frequency of measured daily values (Pfannerstill et al., 2014). The disadvantage of graphical PMs is that model performance can be obtained only qualitatively through them. In addition, graphical PMs can easily be manipulated to look good by scaling.

Table 4 lists a variety of graphical PMs used commonly to support and present results of H/WQ model calibration and validation. The graphical PMs are grouped into two broad categories (direct and derived) to enable users to determine appropriate graphical PMs for their study.

The spatial and temporal scale of simulation could be used to determine graphical performance measures that will be effective in communicating model performance to end users. The most effective graphical measures are ones that highlight specific predictive capabilities of the model. For shorter-term modeling (<1 year), a time series plot can be an effective tool. The performance of models for longer-duration datasets (=10 years

of daily data) is better understood by using either a scatter plot or a duration curve. For instance, when Duda et al. (2012) presented the daily-scale five-year calibration results for an HSPF model application, they provided both a time series graph and a duration curve. The time series graph, which contained approximately 1825 data points, gave the impression that the model sometimes overestimated or underestimated peak flows, depending on the peak. This presented a confusing picture of model performance. The authors then presented the same data in the form of a flow duration curve. The flow duration curve not only indicated that, in general, the model-simulated values were close to the observed values (similar to what was understood from the time series plot), but it also showed that the model overestimated higher flows and underestimated medium and lower flows during the validation period. Thus, the duration curve was a more effective tool for understanding and communicating daily model performance for their case study. The effectiveness of using a duration curve is also demonstrated in a case study presented later in this article.

As discussed in table 4, certain derived graphical PMs, such as cumulative plots and maps, can provide a misleading picture of model performance. For instance, a combination of cumulative and daily time series plot was used by Bottcher et al. (2012) to present results of the WAM model (fig. 1). The presentation of these two plots was essential because the cumulative plot gives the impression that the model overpredicts initially and underpredicts in the latter part of simulation but has reasonable overall performance. On the other hand, the time series plot shows that certain important flow peaks were completely missed. The time series plot allows the modeler to find temporal mismatches that could go unnoticed by using only a cumulative plot.

Maps are also effective tools for presenting key results and meeting the objectives of watershed models. For example, to build confidence in an uncalibrated SWAT model, Srinivasan et al. (2010) used maps to show that SWAT-simulated annual corn and soybean yields for each subbasin were consistent with USDA-NASS estimates. Pai et al. (2011) and Daggupati et al. (2011) used maps of sediment, total P, and nitrate-N outputs to prioritize subwatersheds and fields in SWAT model applications in Arkansas and Kansas. Such maps could be used to assess spatial model performance.

Table 3. Reported performance evaluation criteria for models in the Moriasi et al. (2012) special collection.

Model (and Reference)	Response Output	Performance Evaluation Criteria							
BASINS/HSPF (Duda et al., 2012)		Difference between Simulated and Recorded Values (%)							
	Very Good	Good		Fair					
	Hydrology/flow	<10		10 to 15		15 to 25			
	Sediment	<20		20 to 30		30 to 45			
	Water temperature	<7		8 to 12		13 to 18			
	Water quality/nutrients	<15		15 to 25		25 to 35			
	Pesticides/toxics	<20		20 to 30		30 to 40			
	Hydrology/flow	Statistical Evaluation Criteria							
	Statistic	Very Good		Good		Fair		Poor	
	Daily	R		=0.89 ^[a]		=0.84		=0.77	<0.77
	Monthly	R		=0.92		=0.87		=0.81	<0.81
	Daily	R ²		=0.80		=0.70		=0.60	<0.60
	Monthly	R ²		=0.85		=0.75		=0.65	<0.65

DRAINMOD (Skaggs et al., 2012)			Statistical Evaluation Criteria				
	Statistic		Excellent		Good	Acceptable	
	Water table depth (daily)		MAE (cm)	<10		<15	<20
			NSE	>0.75		>0.60	>0.40
	Drainage volume (cm ³ cm ⁻²)						
	Daily		NSE	>0.75		>0.60	>0.40
	Monthly		NSE	>0.80		>0.70	>0.50
	Annual		NSE	>0.85		>0.75	>0.60
			NPE	<5%		<15%	<25%
EPIC/APEX (Wang et al., 2012)		Satisfactory Calibration Criteria					
		R ²	NSE	PBIAS	Mean and SD	Graphical	
	Runoff or water yield	=0.60	=0.55	Within 20%	-	Simulated time-series flow captures the trend or pattern of measured data.	
	Crop yield	=0.60	-	Within 25%	-	Simulated time-series crop yield captures the trend or pattern of measured data.	
	Sediment yield	=0.60	=0.50	Within 35%	Simulated mean and SD compare closely with measured values	Simulated time-series sediment yield captures the trend or pattern of measured data.	
	Nutrient loss	=0.60	=0.50	Within 50%	-	Simulated time-series nutrient loss captures the trend or pattern of measured data.	
KINEROS/AGWA (Goodrich et al., 2012)	Runoff, erosion, sediments			Acceptable Model Performance			
	Simulated values within 30% of observed (Al-Qurashi et al., 2008)						
RZWQM2 (Ma et al., 2012)	Hydrology, plant growth, nutrients, pesticides			Acceptable Model Simulation			
	R ²		NSE		d	PBIAS	
	=0.80		=0.70		=0.70	Within 15%	
WARMF (Herr and Chen, 2012)				Good Model Performance			
	Hydrology/flow			<20% absolute error			
	Nutrients			<30% absolute error			

Phytoplankton and suspended sediment	<50% absolute error
[a] Values estimated from figure 4 (Duda et al., 2012).	

Statistical Performance Measures

Statistical PMs are widely used to quantify the performance of H/WQ models in describing the closeness of the simulated behavior to observations. Table 5 summarizes commonly used statistical PMs based on the Moriasi et al. (2012) special collection, along with their demonstrated advantages/disadvantages, ranges, optimal values, and the equations used to compute them. Harmel et al. (2014), Bennett et al. (2013), Krause et al. (2005), and Coffey (2004) also provide a comprehensive list of statistical PMs. Although there are different ways to categorize PMs (Moriasi et al., 2007; Bennett et al., 2013), the PMs in this article are grouped as standard regression, dimensionless, and error index, as discussed below.

Standard Regression

Pearson's correlation coefficient (r) and coefficient of determination (R^2) describe the degree of collinearity between simulated and measured data. The correlation coefficient is an index that is used to investigate the degree of linear relationship between observed and simulated data. R^2 is the squared value of r , although it can also be expressed as the squared ratio between the covariance and the multiplied standard deviations of the observed and predicted values (Krause et al., 2005).

Table 4. Summary of graphical performance measures for H/WQ model calibration and validation.

	Purpose	Advantages/Disadvantages
Direct comparison		
Scatter plots	Compare observed and simulated data with no dependent variable. A least square regression line can be fitted to observe deviation from the 1:1 line.	<p>Advantages: Divergence from the 1:1 line provides a visual understanding of the underlying behavior of the model, including any bias or systematic variance.</p> <p>Disadvantages: Data points clumped in the low intensity, high frequency range and few in the high intensity, low frequency range can artificially make a model's performance look good.</p>
Time-series plots	Compare observed and simulated data with time as a dependent variable.	<p>Advantages: Helps inspect and support troubleshooting event-specific prediction issues, including mismatches in magnitude of peaks and shape of recession curve, and outliers. Time series plots can also guide selection of parameters to be used for calibration.</p> <p>Disadvantages: Time series plots become cluttered with too many data points.</p>
Derived comparison		
Cumulative plots	Compare cumulative observed and simulated values with time as	<p>Advantages: Allows identification of any systematic temporal divergence between observed and simulated values.</p>

	dependent variable.	Disadvantages: Cumulative plots may still converge, with major temporal mismatches. They should be used as a preliminary model performance-screening tool.
Flow and load duration curves	Compare observed and simulated values with probability as a dependent variable.	Advantages: Provides insight into model performance over different flow/load regimes (i.e., low, medium, high; Pfannerstill et al., 2014). Disadvantages: Needs a larger number of data points to derive meaningful conclusions. Duration curves are most useful for long-term monthly, daily, or subdaily calibrations.
Maps	Map showing the output of interest at the desired spatial scale. Examples include showing annual sediment loss for each subwatershed.	Advantages: Useful for presenting field-scale to watershed-scale model results for understanding the spatial performance of the model. Pollutant hotspots within a watershed can be quickly identified using color-codes. Disadvantages: Choices of color-coding and grouping within a map can sometimes be misleading. For example, red colored areas may or may not represent critical areas depending on actual values plotted.

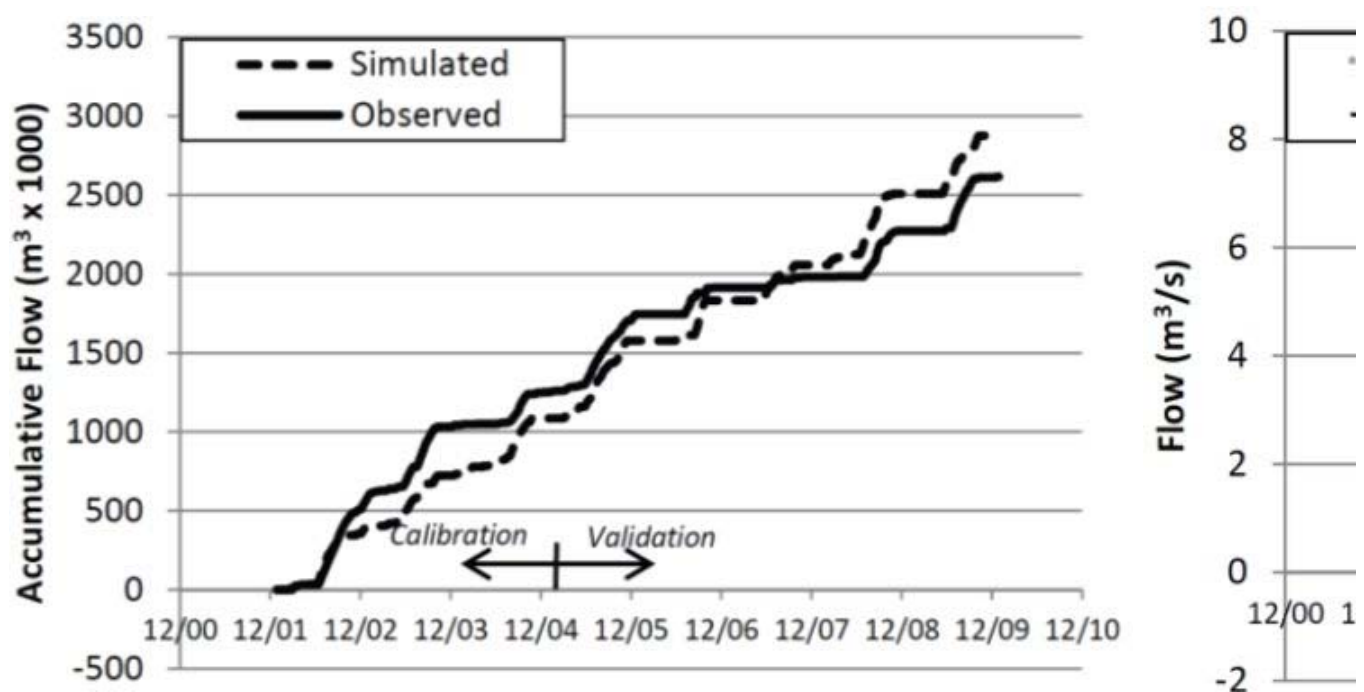


Figure 1. Calibrated daily flow using the WAM model (reproduced from Bottcher et al., 2012).

Dimensionless

The Nash-Sutcliffe efficiency (NSE) is a normalized statistic that determines the relative magnitude of the residual variance (noise) compared to the measured data variance (information; Nash and Sutcliffe, 1970). NSE indicates how well the plot of observed versus simulated data fits the 1:1 line. Many studies (e.g., Santhi et al., 2001; Vazquez-Amabile and Engel, 2005; Reungsang et al., 2010; Pai et al., 2011; Douglas-Mankin et al., 2013) have used NSE to evaluate model performances for various output responses (e.g., flow, sediment, N, P, crop yields, etc.) using different models (MIKE-SHE, ADAPT, SWAT, WARMF, HSPF, etc.).

The index of agreement (d) was developed by Willmott (1981) as a standardized measure of the degree of model prediction error. The index of agreement represents the ratio between the mean square error and the potential error (Willmott, 1984). The potential error (denominator in index of agreement equation in

table 5) represents the largest value that the squared difference of each pair can attain. The index of agreement can detect additive and proportional differences in the observed and simulated means and variances.

Error Index

The root mean square error (RMSE) is the square root of mean square error (MSE). The MSE is also known as standard error of the estimate in regression analysis. The RMSE is measured in the same units as the model output response of interest and is representative of the size of a typical error.

Table 5. Equations, ranges, optimal values, and advantages and disadvantages for statistical performance measures in the Moriasi et al. (2012) special collection (O and P are observed and predicted values, respectively).

Statistic	Equation	Range	Optimal Value	Advantages/Disadvantages
r	$\frac{\sum_{i=1}^n (O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^n (O_i - \bar{O})^2} \sqrt{\sum_{i=1}^n (P_i - \bar{P})^2}}$	-1.0 to 1.0	-1.0 (negative slope) or 1.0 (positive slope)	<p>Advantages: R^2 and r are widely used in hydrological modeling studies, thus serving as a benchmark for performance evaluation.</p> <p>Disadvantages: R^2 and r are oversensitive to high extreme values (Krause et al., 2005) and insensitive to additive and proportional differences between model predictions and measured data (Legates and McCabe, 1999).</p> <p>Notes: We recommend that the regression line gradient and intercept be reported when R^2 is used as a performance measure. For a good agreement, the intercept should be close to zero and the gradient should be close to one (Krause et al., 2005).</p>
R^2	$\left[\frac{\sum_{i=1}^n (O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^n (O_i - \bar{O})^2} \sqrt{\sum_{i=1}^n (P_i - \bar{P})^2}} \right]^2$	0.0 to 1.0	1.0	<p>Advantages: NSE is: (1) a quantitative measure conducive to development of PEC; (2) good for use with continuous long-term simulations and can be used to determine how well the model simulates trends for the output response of concern; (3) robust and can be used to evaluate model performance for several output responses (e.g., stream flow, sediments, nutrients, pesticides) and temporal scales; and (4) commonly used, which means that there is extensive information on reported values, which can be used for comparison purposes. Further, it can incorporate measurement uncertainty (Harmel and Smith, 2007; Harmel et al., 2010).</p> <p>Disadvantages: NSE cannot help identify model bias and cannot be used to identify differences in timing and magnitude of peak flows and shape of recession curves; in other words, it cannot be used for</p>
NSE	$1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2}$	-8 to 1.0	1.0	

single-event simulations.

Notes: NSE is sensitive to extreme values due to the squared differences (Krause et al., 2005). To overcome extreme-value cases and increase sensitivity to lower measured and simulated values, Krause et al. (2005) recommended the use of logarithmic and relative derivatives forms of NSE and d . In cases where the measured data are bi-modal with high and low distributions in the same study area, such as the measured flows in Cho and Olivera (2009), it is recommended that the two data categories be separated to avoid the bias toward simulation of lower values.

				<p>Advantages: The index of agreement (1) detects additive and proportional differences in the observed and simulated means and variances and (2) is widely used, and thus there is comprehensive information on reported values in the literature.</p> <p>Disadvantages: Overly sensitive to extreme values due to the squared differences (Legates and McCabe, 1999). High values of d were reported even for poor model fits (Krause et al., 2005).</p> <p>Notes: d should be evaluated based on the phenomenon studied, measurement accuracy, and the model employed. It can also be used as a substitute for R^2 to identify the degree to which model predictions are error-free (Legates and McCabe, 1999). Further, it can incorporate measurement uncertainty (Harmel and Smith, 2007; Harmel et al., 2010).</p>
d	$1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (P_i - \bar{O} + O_i - \bar{O})^2}$	0.0 to 1.0	1.0	
				<p>Advantages: RMSE and MAE are: (1) computed and reported in the same units as the model output of concern and are hence easy for readers to interpret; (2) work well for continuous long-term simulations; and (3) commonly used in model performance evaluation.</p> <p>Disadvantages: Error indices are measured in the same unit as the model output being investigated, so they cannot be used by themselves to gauge model performance for values other than zero.</p> <p>Notes: RMSE and MAE can be used to determine confidence intervals in model predictions, and it is possible to incorporate measurement uncertainty (Harmel and</p>
RMSE or RMSD	$\sqrt{\frac{1}{n} \sum_{i=1}^n (O_i - P_i)^2}$	0.0 to 8	0.0	
MAE	$\frac{1}{n} \sum_{i=1}^n O_i - P_i $	0.0 to 8	0.0	

				Smith, 2007; Harmel et al., 2010).
RSR	$\frac{\sqrt{\sum_{i=1}^n (Q_i - P_i)^2}}{\sqrt{\sum_{i=1}^n (Q_i - \bar{P})^2}}$	0.0 to 8	0.0	<p>Advantages: RSR incorporates the benefits of error index statistics and includes a scaling/normalization factor, so the resulting statistics and reported values can apply to various output responses.</p> <p>Disadvantages: RSR gives more weight to high values when compared with low values because errors in high values are usually greater in absolute value than errors in low values due to the squared difference values in the denominator.</p> <p>Notes: RSR has not been widely used in the H/WQ modeling literature since it is a relatively new statistical performance measure.</p>

Table 5 (continued). Equations, ranges, optimal values, and advantages and disadvantages for statistical performance measures in the Moriasi et al. (2012) special collection (*O* and *P* are observed and predicted values, respectively).

Statistic	Equation	Range	Optimal Value	Advantages/Disadvantages
RE or PE	$\frac{ Q_i - P_i }{Q_i} \times 100$	0.0 to 8 to 8	0.0	<p>Advantages: (1) RE facilitates comparison of model performance between different output responses, and (2) differences between observed and predicted values are quantified as relative deviations. This significantly reduces the influence of absolute differences during high flows.</p> <p>Disadvantages: The absolute lower differences during low flow periods are enhanced because they are significant if looked at in a relative sense. As a result, there might be a systematic over- or underprediction during low flow periods.</p> <p>Notes: RE can be used along with other statistics to quantify low flow simulations</p>
PBIAS	$\frac{\sum_{i=1}^n Q_i - P_i}{\sum_{i=1}^n Q_i} \times 100$	-8 to 8	0.0	<p>Advantages: PBIAS: (1) can be used to determine how well the model simulates the average magnitudes for the output response of interest; (2) is useful for continuous long-term simulations; (3) is robust and commonly used, which means that there is extensive information on reported values; (4) can help identify average model simulation bias (overprediction vs. underprediction); and (5) can incorporate measurement uncertainty (Harmel et al., 2010).</p> <p>Disadvantages: PBIAS cannot be used (1)</p>

for single-event simulations to identify differences in timing and magnitude of peak flows and the shape of recession curves nor (2) to determine how well the model simulates residual variations and/or trends for the output response of interest.

Notes: PBIAS can give a deceiving rating of model performance if the model overpredicts as much as it underpredicts, in which case PBIAS will be close to zero even though the model simulation is poor. It is therefore recommended that PBIAS be used with other statistical and graphical PMs to determine model performance.

The mean absolute error (MAE) is also measured in the same units as the model output response of interest. It is usually similar in magnitude but slightly smaller than the RMSE. The RMSE also tends to give more weight to high values than low values because errors in high values are usually greater in absolute value than errors in low values (Gan et al., 1997; Gan and Biftu, 1996; Eckhardt and Arnold, 2001; van Griensven and Bauwens, 2003; Huisman et al., 2003; Cho and Olivera, 2009). To get around this limitation, Moriasi et al. (2007) recommended that RMSE be normalized using the observations standard deviation, giving a measure referred to as the RMSE-observations standard deviation ratio (RSR).

Although it is commonly accepted that the lower the RMSE, the better the model performance, only Singh et al. (2004) published a guideline to qualify what is considered a low RMSE based on the observations standard deviation (SD). Singh et al. (2004) stated that RMSE values of less than half of the SD of the observations may be considered low. Based on the recommendation by Singh et al. (2004), Moriasi et al. (2007) developed the RSR.

Relative error (RE), absolute relative error, or absolute relative deviation is the ratio of absolute error of the simulated data to the observed data. It indicates the mismatch that occurs between the observed and modeled values, expressed in terms of ratios and percentages. Krause et al. (2005) recommended relative efficiency criteria for NSE and d in which relative deviations are derived for NSE and d . These can be used to quantify low flow simulations. Relative bias (RB), relative volume error (RVE), and many other bias-based statistics are derived based on RE to report statistical PMs in evaluating hydrological model performances.

Percent bias (PBIAS) measures the average tendency of the simulated data to be larger or smaller than observed counterparts (Gupta et al., 1999). It also measures over- and underestimation of bias and expresses it as a percentage. Percent stream flow volume error (PVE; Singh et al., 2004), prediction error (PE; Fernandez et al., 2005), and percent deviation of stream flow volume (D_v ; ASCE, 1993; Moriasi et al., 2007) are calculated in a similar manner as PBIAS.

Meta-Analysis of Performance Data

Reported Value Ranges for Performance Measures

For each model included in the Moriasi et al. (2012) special collection, approximately 20 available peer-reviewed articles were collected. Performance data for case studies in the Moriasi et al. (2012) special collection and for articles reviewed by Moriasi et al. (2007) were not considered in this study. While this effort was by no means exhaustive, it yielded a sizeable dataset including 312 data points for R^2 and 435 data points for NSE that were used in the meta-analysis. Due to the volume of material involved, reported performance data for each simulated component during calibration and validation were recorded

(supplemental material tables S1-1 through S1-22, available at http://bit.ly/NRES_SW10715). These data were collected from articles published from 1992 to 2013; 93% were published in 2000 or later, and 53% were published after 2007. Most of the reported parameters are for field-scale (tables S1-2 to S1-10) and watershed-scale (tables S1-11 to S1-22) models that utilize both manual and autocalibration methods. Of the reviewed articles, most reported model calibration and validation on flow-related components (tables S1-2 to S1-5 and S1-11 to S1-15), and most are based on the SWAT model. The least reported model calibration and validation PM values were those associated with point to plot scale models (table S1-1). Most of the models in this category utilize autocalibration algorithms that select all possible combinations of solutions that meet the set threshold for the selected objective function.

Of the models examined (table 1), only SWAT, HSPF, WARMF (watershed-scale), and ADAPT (field-scale) had sufficient model performance data for meaningful analyses. The total numbers of reviewed articles from which data were obtained for analyses of SWAT, HSPF, WARMF, and ADAPT models were 33, 17, 2, and 16, respectively. For each of the aforementioned models, values for R^2 , NSE, and PBIAS were reported most frequently, but there was also an appreciable amount of data on the index of agreement (d) at field scale. Based on reviewed literature, point to plot (and to some extent field-scale) models used different simulated response outputs to evaluate model performance. For instance, Essaid et al. (2008) and Healy and Essaid (2012) used streambed water flux and temperature to evaluate VS2DI performance, while Huth et al. (2012) used soil water content to evaluate SWIM3. Krobek et al. (2010) and Diekkruiger et al. (1995) also used soil water content to evaluate the performance of the DAISY model. The use of different simulated response outputs and the limited amount of reported peer-reviewed model performance data made it difficult to conduct statistical comparisons for these smaller spatial scale models, so they were excluded from the analysis and PEC development.

Preliminary Diagnostics of Data Used for Meta-Analysis

Table 6 summarizes the data used for the meta-analysis. Based on a preliminary analysis, reported performance data values for watershed-scale models, irrespective of output response and temporal scale, varied from 0.02 to 1.00 for R^2 , from -10.30 to 0.99 for NSE, and from -81.1% to 167% for PBIAS (table 4). Reported R^2 values for field-scale models for flow at a monthly temporal scale varied from 0.18 to 0.91, while d values varied from 0.60 to 0.99 (table 6).

Table 6. Summary of the performance data used for detailed statistical analyses.						
Performance Measure			Temporal Scale ^[a]			
			Annual	Monthly	Daily	Seasonal
Watershed scale						
	R^2	Entries	89	196	27	-
		Mean	0.67	0.63	0.63	-
		Median	0.67	0.72	0.70	-
		Minimum	0.32	0.18	0.02	-
		Maximum	1.00	0.99	0.97	-
	NSE	Entries	87	233	115	-
		Mean	0.58	0.44	0.13	-
		Median	0.60	0.59	0.53	-
		Minimum	-0.91	-7.89	-10.3	-

		Maximum	0.99	0.96	0.96	-
	PBIAS	Entries	26	57	-	29
		Mean	-14.92	7.51	-	20.4
		Median	0	6.4	-	8
		Minimum	-81.1	-38.4	-	-46.4
		Maximum	35.3	53.1	-	167
Field scale						
	R^2	Entries	-	29	-	-
		Mean	-	0.74	-	-
		Median	-	0.75	-	-
		Minimum	-	0.18	-	-
		Maximum	-	0.91	-	-
	d	Entries	-	33	-	-
		Mean	-	0.88	-	-
		Median	-	0.91	-	-
		Minimum	-	0.60	-	-
		Maximum	-	0.99	-	-
<p>[a] ❖❖❖ Blank entries mean that either there were no data or that available data were insufficient for meaningful statistical analyses. All available raw data are presented in the supplemental material tables (available at http://bit.ly/NRES_SW10715).</p>						

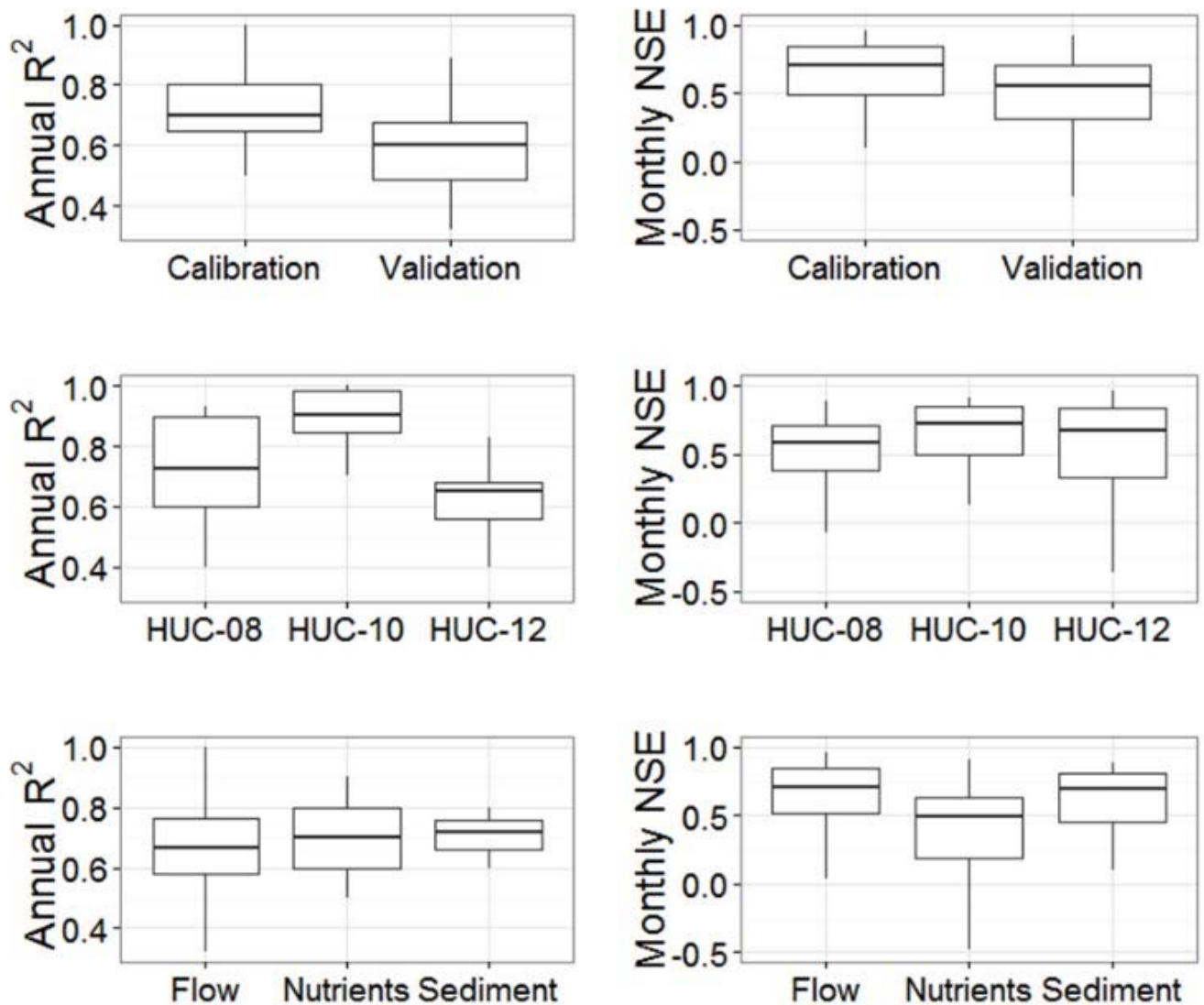


Figure 2. Box and whisker plots showing comparisons of performance data considering: (top row) calibration and validation data for watersheds at HUC 8 and larger), and (bottom row) simulated component.

Further analysis of the distributions of the combined datasets (regardless of whether they pertained to calibration or validation, watershed size, and/or the components) showed that most tended to be skewed toward the higher values of the specific PMs (table 6 and fig. 2). This was expected

because calibration and validation efforts are usually geared toward finding the best suitable values, which are the highest values for measures such as R^2 , NSE, and d . Exceptions to this trend were values of PBIAS, which were more centrally located. Again, this is not surprising, as PBIAS can vary between small and large values, both negative and positive, and by definition PBIAS values close to zero indicate better model performance and are thus more desirable. The other exception was R^2 values, for which the data were approximately normally distributed. At this point, it is unclear why this was the case. Based on the approximate distributions of the performance data, the nonparametric median test was used to test whether there were significant differences among reported performance values data (table 7) among the various categories to warrant development of separate PEC.

Table 7. Summary of results of the statistical analyses on the performance data.

Comparisons			Temporal Scale and Performance Measure							
			Annual				Monthly			Daily
			R ²	NSE	PBIAS		R ²	NSE	PBIAS	NSE
Watershed scale										
	Calibration vs. validation									
		Calibration entries	57	53	8		106	127	27	66
		Validation entries	32	34	18		90	106	30	49
		p-value ^[a]	0.0047*	0.0112*	0.0401*		0.5674	0.0131*	0.0249*	<0.0001*
	Comparison by HUC									
		HUC-08+ entries	26	4	10		138	118	56	5
		HUC-10 entries	7	6	16		14	54	1	62
		HUC-12 entries	56	76	0		44	61	0	40
		p-value	0.0002*	-	0.0123*		<0.0001*	0.2330	-	0.0158*
	Comparison by component									
		Flow entries ^[b]	84	72	26		88	119	32	88
		Sediment entries	3	4	0		46	31	15	3
		N entries	2	0	0		31	49	10	18
		P entries	0	11	0		31	34		6
		p-value	-	0.0453*	-		<0.0004*	<0.0001*	0.1281	<0.0001*
Field scale							R ²	<i>d</i>		
	Calibration entries						17	18		
	Validation entries						12	15		
	p-value						0.5799	0.3499		

[a] Probability that observed differences in reported performance data values are attributable to error or chance given an a level of significance ($\alpha = 0.05$ in this case). Values α indicate that the reported performance data values (e.g., for calibration vs. validation) are significantly different at that level of significance, with smaller values indicating higher significance (i.e., probability that observed differences were due to error or chance is very small). Asterisks (*) indicate significant

differences in performance data values for calibration vs. validation, HUC, and modeled component.

Combines data for stream flow, surface runoff, and base flow as reported.

For most of the watershed-scale analyses performance data, values for calibration were significantly different (table 7) from those reported for validation, with those for calibration being better (fig. 2). This was not the case for the field-scale data, for which the performance data values were not significantly different between the calibration and validation periods. Ideally, performance values obtained for validation need to be close to those obtained during calibration; a discrepancy between these values is evidence of model divergence (Sorooshian and Gupta, 1995; Duda et al., 2012; Zheng et al., 2012), suggesting calibrated model inaccuracies in process representation (Sorooshian, 1983). Since calibration efforts rely on comparisons between observed and measured data, it is possible to make parameter adjustments simply to suit this kind of comparison while ignoring the accuracy of the process simulation. Thus, in recommending guidelines, we do not make a distinction between calibration and validation periods.

Significant differences were also observed in reported performance data values at the watershed scale, with the exception of monthly NSE values (table 7 and fig. 2). Although no clear patterns were discernible, the models seemed to perform better for HUC-10 watersheds than for HUC-08+ and HUC-12 watersheds. Similarly, at each temporal scale, there were significant differences among PMs based on the response output being simulated and the available data for reported model PM values (table 7). For example, data analysis indicated better simulation of flow than all other response outputs. This was expected, given that hydrologic processes are the primary drivers within a watershed; thus, associated simulated response outputs are calibrated first and more extensively. In addition, more observed data are available to calibrate models for flow than for sediments or nutrients.

Further analyses based on both simulated response output and temporal scale (e.g., annual flow, monthly flow, etc.) also showed significant differences for R^2 and NSE ($p = 0.0002$ and 0.0001 , respectively), although no significant differences were observed among the temporal scales when all data were grouped together and analyzed solely by temporal scale ($p = 0.0661$, 0.1957 , and 0.0811 for R^2 , NSE, and PBIAS, respectively). Due to the difficulties in duplicating the timing of flow, and given the uncertainties in the timing of model inputs (mainly precipitation; Duda et al., 2012), model calibration is considered to be simpler at the annual temporal scale and is progressively more difficult as the temporal scale resolutions becomes finer (Engel et al., 2007; Moriasi et al., 2007; Duda et al., 2012). Thus, this latter finding was somewhat surprising. However, the art of model calibration has greatly improved in recent years due to model autocalibration tools and techniques. These are designed to find optimal parameters based on PMs, hence increasing the likelihood that resulting model PM values will be comparable regardless of the temporal scale.

Based on the meta-analysis results, we determined that there was a need for separate PEC for each of the commonly simulated response outputs, watershed- and field-scale models, temporal scales, and for the recommended PMs. However, there was also the need for general PEC that could be used across temporal scales. The final recommended PEC for the identified separate categories are based primarily on the results of computed percentiles of reported performance data to determine thresholds for the different qualitative ratings used in this article, existing PEC (Al-Qurashi et al., 2008; Moriasi et al., 2007; Duda et al., 2012; Herr and Chen, 2012; Ma et al., 2012; Skaggs et al., 2012; Wang et al., 2012), and our modeling experience.

Development of Criteria for Selected Statistical Performance Measures

The final step of the meta-analysis was to compute percentiles of available performance data to develop separate PEC for R^2 , NSE, PBIAS, and d for the spatial and temporal scales and simulated response outputs identified by the median test in the previous subsection. There were 57 negative NSE values reported for watershed-scale models (supplemental material tables S1-11 to S1-20). However, by definition, $NSE < 0.0$

indicates that the mean observed value is a better predictor than the simulated value, which indicates unacceptable performance. Therefore, all negative values for NSE were excluded. While we agree that NSE is more stringent than R^2 or d , we did not exclude any reported performance data for R^2 and d corresponding to the studies that reported negative NSE. This is because different PMs have varied strengths that aid in determining the performance of a given model during the calibration and validation periods. Therefore, the reported performance data for each PM were analyzed independently.

To be consistent with model PEC previously recommended by Moriasi et al. (2007), \diamond very good, \diamond good, \diamond satisfactory, \diamond and \diamond not satisfactory \diamond ratings were defined. Initial PEC were then developed for each of the ratings based on different data distributions at spatial and temporal scales and simulated response outputs for the recommended criteria. Even though percentile is used to measure spread, we also found it appropriate to use as an initial step in determining the thresholds for the defined ratings due to the fact that the calibration process seeks to optimize PMs for response outputs of interest. Considering the ranges of model PM data obtained (table 6) and expected reasonable PM data values, model performance values at and below the 25th percentile were considered \diamond not satisfactory, \diamond model performance values between the 25th to 50th percentiles were considered \diamond satisfactory, \diamond model performance values within and including the 50th to 75th percentiles were considered \diamond good, \diamond and those above the 75th percentile were considered \diamond very good. \diamond Values obtained based on percentiles were adjusted accordingly (e.g., rounded off) to produce meaningful intervals for these initial PEC. Figure 3 shows an example of the PEC development process. To facilitate PEC development for PBIAS, all related entries were converted into absolute values (fig. 3b). Because of the nature of this statistic, the rating and corresponding percentile ranges were reversed.

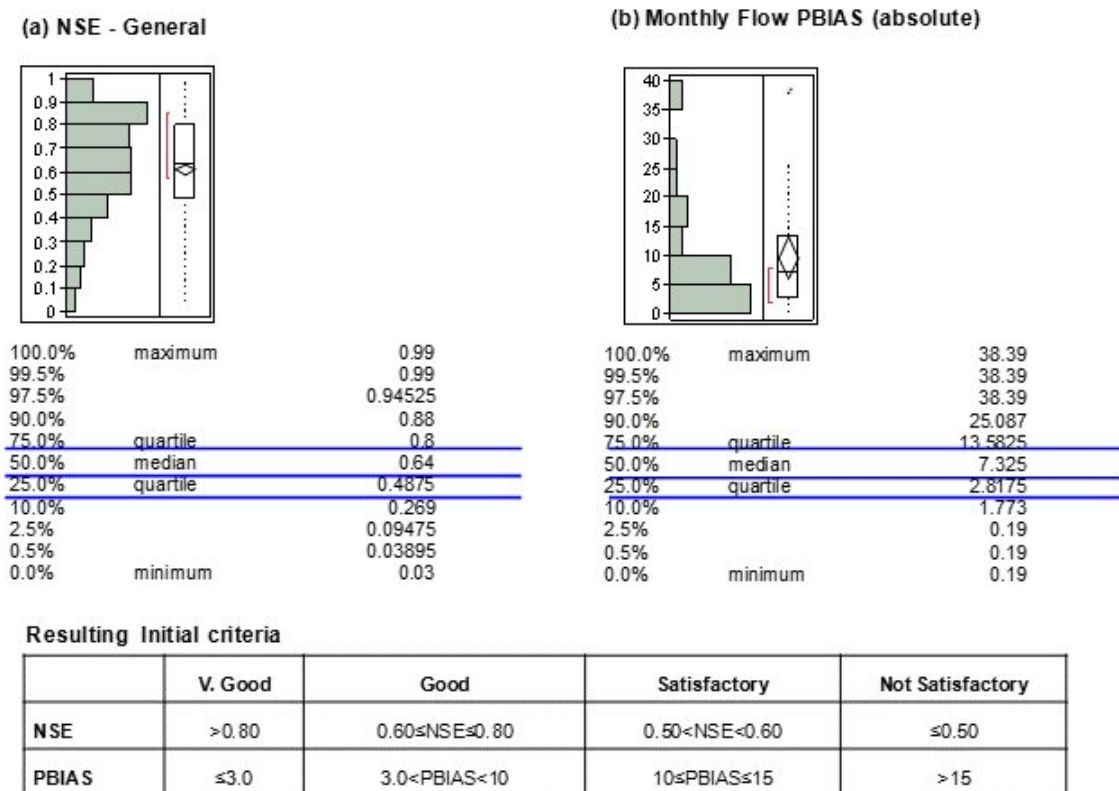


Figure 3. Example of initial performance evaluation criteria development for flow: (a) annual NSE and (b) monthly PBIAS.

Analysis of the initial PEC based on data distributions resulted in several noteworthy differences (table 8). For example, with NSE, the resulting PEC for flow were different from those for N and P, with the former PEC being stricter. This was expected due to the large amount of observed flow calibration data, which is not

the case for sediment and nutrient data. It is also critical that flow simulation be accurate, as flow is the primary driver of watershed processes. Sediment seemed to exhibit a similar response to flow, possibly for the same reasons. This explains why PEC were stricter for flow than for N and P.

With regard to temporal scale, however, the distinctions were not as clear. While data were not always sufficient to allow comparisons for each component, in some instances the resulting PEC were contradictory, e.g., initial PEC were stricter for monthly flow than for annual flow. This was in contrast to Moriasi et al. (2007), who suggested more relaxed PEC for a daily temporal scale and progressively higher thresholds for subsequent coarser temporal scales. As previously discussed, our data did not show significant differences on the basis of temporal scale alone, which could possibly explain these discrepancies. For each of the PMs, general initial PEC (table 8) were also derived independent of either component or temporal scale and seemed to offer more unifying values that could be used as alternates where contradictions were encountered.

As a final step, the initial PEC were reviewed and revised based on previous PEC as reported in the literature (Al-Qurashi et al., 2008; Moriasi et al., 2007; Duda et al., 2012; Herr and Chen, 2012; Ma et al., 2012; Skaggs et al., 2012; Wang et al., 2012) and on our modeling experience. The final PEC developed are reported under the ♦Guidelines for Model Performance Evaluation: Recommended Measures and Criteria♦ subheading.

Table 8. Initial performance evaluation criteria for recommended statistical performance measures for watershed- and field-scale models based on the distribution of existing data.								
Measure		Component	Temporal Scale	<i>n</i>	Very Good	Good	Satisfactory	Not Satisfactory
Watershed scale								
	R^2	Flow	Annual	84	>0.75	$0.70 = R^2 = 0.75$	$0.60 < R^2 < 0.70$	$=0.60$
			Monthly	87	>0.85	$0.80 = R^2 = 0.85$	$0.70 < R^2 < 0.80$	$=0.70$
			Daily	27	>0.85	$0.70 = R^2 = 0.85$	$0.50 < R^2 < 0.70$	$=0.50$
		Sediment	Annual	3	-	-	-	-
			Monthly	46	>0.80	$0.65 = R^2 = 0.80$	$0.40 < R^2 < 0.65$	$=0.40$
			Daily	0	-	-	-	-
		N	Annual	2	-	-	-	-
			Monthly	31	>0.70	$0.60 = R^2 = 0.70$	$0.30 < R^2 < 0.60$	$=0.30$
			Daily	0	-	-	-	-
		P	Annual	0	-	-	-	-
			Monthly	31	>0.80	$0.65 = R^2 = 0.80$	$0.40 < R^2 < 0.65$	$=0.40$
			Daily	0	-	-	-	-

		General		311	>0.80	$0.70 = R^2 = 0.80$	$0.50 < R^2 < 0.70$	=0.50
	NSE	Flow	Annual	71	>0.75	$0.60 = NSE = 0.75$	$0.50 < NSE < 0.60$	=0.50
			Monthly	109	>0.85	$0.70 = NSE = 0.85$	$0.55 < NSE < 0.70$	=0.55
			Daily	79	>0.80	$0.70 = NSE = 0.80$	$0.50 < NSE < 0.70$	=0.50
		Sediment	Annual	4	-	-	-	-
			Monthly	31	>0.80	$0.70 = NSE = 0.80$	$0.45 < NSE < 0.70$	=0.45
			Daily	3	-	-	-	-
		N	Annual	0	-	-	-	-
			Monthly	31	>0.70	$0.60 = NSE = 0.70$	$0.35 < NSE < 0.60$	=0.35
			Daily	6	>0.55	$0.40 = NSE = 0.55$	$0.25 < NSE < 0.40$	=0.25
		P	Annual	10	>0.65	$0.60 = NSE = 0.65$	$0.50 < NSE < 0.60$	=0.50
			Monthly	33	>0.65	$0.50 = NSE = 0.65$	$0.40 < NSE < 0.50$	=0.40
			Daily	1	-	-	-	-
		General		378	>0.80	$0.60 = NSE = 0.80$	$0.50 < NSE < 0.60$	=0.50
	PBIAS (%) ^[a]	Flow	Annual	26	= $\diamond 2.5$	$\diamond 2.5 < \text{PBIAS} < \diamond 15$	$\diamond 15 = \text{PBIAS} = \diamond 35$	> $\diamond 35$
			Monthly	32	= $\diamond 3.0$	$\diamond 3.0 < \text{PBIAS} < \diamond 10$	$\diamond 10 = \text{PBIAS} = \diamond 15$	> $\diamond 15$
			Seasonal	29	= $\diamond 10$	$\diamond 10 < \text{PBIAS} < \diamond 15$	$\diamond 15 = \text{PBIAS} = \diamond 45$	> $\diamond 45$
		Sediment	Annual	0	-	-	-	-
			Monthly	15	= $\diamond 1$	$\diamond 1 < \text{PBIAS} < \diamond 10$	$\diamond 10 = \text{PBIAS} = \diamond 20$	> $\diamond 20$
			Seasonal	0	-	-	-	-
		Nutrients	Annual	0	-	-	-	-

			Monthly	10	$=\diamond 10$	$\diamond 10 < \text{PBIAS} < \diamond 15$	$\diamond 15 = \text{PBIAS} = \diamond 30$	$>\diamond 30$
			Seasonal	0	-	-	-	-
		General		112	$=\diamond 5$	$\diamond 5 < \text{PBIAS} < \diamond 10$	$\diamond 10 = \text{PBIAS} = \diamond 25$	$>\diamond 25$
Field scale								
	R^2		Monthly	29	>0.85	$0.75 = R^2 = 0.85$	$0.70 < R^2 < 0.75$	$=0.70$
	d		Monthly	33	>0.90	$0.85 = d = 0.90$	$0.75 < d < 0.85$	$=0.75$
[a] $\diamond\diamond\diamond$ Values are absolute.								

Guidelines for Model Performance Evaluation: Recommended Measures and Criteria

Prior to providing any general recommendations for model PMs and their corresponding PEC, we note that it is critical that model users follow proper calibration and validation procedures to obtain the correct model performance for the right reasons (Kirchner, 2006; Arnold et al., 2015). In this regard, we recommend that model users should consider recommendations for all other key calibration and validation topics covered in this special collection. These include (1) ensuring that terminology is clearly defined (Zeckoski et al., 2015), (2) selecting an appropriate model based on the study goals and ensuring that the model and fluxes are well represented (Arnold et al., 2015), (3) considering appropriate spatial and temporal scales (Baffaut et al., 2015), (4) parameterizing the model appropriately (Malone et al., 2015), and (5) employing appropriate calibration and validation strategies (Daggupati et al., 2015b), including sensitivity (Yuan et al., 2015) and uncertainty (Guzman et al., 2015) analyses. Having taken all these important modeling aspects into consideration, model users should then use appropriate PMs along with the corresponding general PEC recommended in this article. Finally, we recommend that all these aspects of modeling be properly documented and reported (Saraswat et al., 2015) with sufficient detail to ensure repeatability.

The first step in evaluating model performance is to use recommended graphical PMs because they provide a visual indication of model performance. The next step is to compute values for the recommended statistical PMs. The computed values are then compared with recommended PEC to assess model performance with respect to statistical PMs.

Table 9. Final performance evaluation criteria for recommended statistical performance measures for watershed- and field-scale models.

Measure	Output Response	Temporal Scale ^[a]	Performance Evaluation Criteria			
			Very Good	Good	Satisfactory	Not Satisfactory
Watershed scale						

	R ²	Flow ^[b]	D-M-A	R ² > 0.85	0.75 < R ² = 0.85	0.60 < R ² = 0.75	R ² = 0.60
		Sediment/P ^[c]	M	R ² > 0.80	0.65 < R ² = 0.80	0.40 < R ² = 0.65	R ² = 0.40
		N	M	R ² > 0.70	0.60 < R ² = 0.70	0.30 < R ² = 0.60	R ² = 0.30
	NSE	Flow	D-M-A	NSE > 0.80	0.70 < NSE = 0.80	0.50 < NSE = 0.70	NSE = 0.50
		Sediment	M	NSE > 0.80	0.70 < NSE = 0.80	0.45 < NSE = 0.70	NSE = 0.45
		N/P ^[c]	M	NSE > 0.65	0.50 < NSE = 0.65	0.35 < NSE = 0.50	NSE = 0.35
	PBIAS (%)	Flow	D-M-A	PBIAS < 5	5 = PBIAS < 10	10 = PBIAS < 15	PBIAS = 15
		Sediment	D-M-A	PBIAS < 10	10 = PBIAS < 15	15 = PBIAS < 20	PBIAS = 20
		N/P ^[c]	D-M-A	PBIAS < 15	15 = PBIAS < 20	20 = PBIAS < 30	PBIAS = 30
Field scale							
	R ²	Flow	M	R ² > 0.85	0.75 < R ² = 0.85	0.70 < R ² < 0.75	R ² = 0.70
	d	Flow	M	d > 0.90	0.85 < d = 0.90	0.75 < d < 0.85	d = 0.75
<p>[a] D, M, and A denote daily, monthly, and annual temporal scales, respectively.</p> <p>[b] Includes stream flow, surface runoff, base flow, and tile flow, as appropriate, for watershed- and field-scale models.</p> <p>[c] Where there were no differences, PEC were grouped for the output responses.</p>							

Recommended Performance Measures

Due to varied strengths of the different PMs described in this article, we recommend the use of multiple graphical and statistical PMs. Both direct and derived graphical PMs are recommended in determining model calibration and validation performance. For shorter periods and coarse temporal resolutions (e.g., monthly calibration for one to three years), time series and scatter plots are most effective for data visualization and demonstration of model performance. With increasing data points, an inconsistent understanding of model performance may result from direct graphical PMs. Under such circumstances, derived measures such as cumulative distributions or duration curves should be employed. For field- and watershed-scale models, where calibration and validation are done at the outlet, we recommend using maps to ensure that non-calibrated locations provide reasonable values for outputs of interest such as soil erosion or nutrient loss. This will ensure a more comprehensive evaluation of model performance and confidence in model outputs.

The most commonly used statistical PMs with varied complementary strengths are recommended. These include R^2 (in conjunction with the gradient b and the intercept a of the corresponding regression line), NSE, d , RMSE alongside RSR, and PBIAS. These statistics can be used for daily, monthly, and yearly temporal scales and for all major output responses. During low flow simulations, logarithmic or relative derivatives of NSE or d need to be used, as recommended by Krause et al. (2005). We also recommend that RSR be reported alongside RMSE, with RMSE providing model performance in the units of the output response of interest and RSR providing a normalized value for comparison of model performance for various studies.

Recommended Performance Criteria

The recommended PEC for the statistical PMs NSE, R^2 , d , and PBIAS for different output responses at different spatial and temporal scales are presented in table 9. The PEC in table 9 result from a combination of previous PEC as reported in the literature (Al-Qurashi et al., 2008; Moriasi et al., 2007; Duda et al., 2012; Herr and Chen, 2012; Ma et al., 2012; Skaggs et al., 2012; Wang et al., 2012), meta-analysis conducted in this study, and our modeling experience. For a given study, the same PBIAS PEC are recommended for the three temporal scales because PBIAS is computed based on observed daily, monthly, and annual values derived from data collected or measured at a finer temporal scale, such as hourly or sub-hourly. These PEC apply to both model calibration and validation periods. For example, based on table 9, model performance can be judged as \blacklozenge satisfactory \blacklozenge for flow simulations if monthly $R^2 \blacklozenge > 0.70$ and $d > 0.75$ for field-scale models and daily, monthly, or annual $R^2 > 0.60$, $NSE > 0.50$, and $PBIAS = \blacklozenge 15\%$ for watershed-scale models. Although we recommend RMSE (with RSR) and the logarithmic or relative derivative of d or NSE statistical PMs, no PEC were developed for them because the available data were not sufficient for meta-analysis and thus for PEC development. However, for RSR, we recommend that the PEC proposed by Moriasi et al. (2007) be used until new PEC can be developed. The intent of this study was to develop generalizable PEC for all models. However, sufficient data for meta-analysis were available only for SWAT, HSPF, WARMF, and ADAPT, as mentioned earlier. Therefore, we also recommend that the PEC developed in this study be used primarily for these models and used only with caution for other models. For example, in the absence of spatial-specific model criteria, the stated watershed PMs and corresponding criteria can be adopted and/or modified for other spatial scale models.

The PEC recommended in this study are general and can be adjusted as appropriate. However, we consider some values of the recommended PMs to be unacceptable beyond certain reasonable ranges. For example, as explained earlier, we consider negative values of NSE to indicate unacceptable model performance. Unacceptable values of PBIAS can be derived from Harmel et al. (2006), with maximum measurement uncertainties under typical measurement scenarios considered to be $\blacklozenge 19\%$ for stream flow, $\blacklozenge 69\%$ for nitrate-N ($\text{NO}_3\text{-N}$), $\blacklozenge 100\%$ for ammonium-N ($\text{NH}_4\text{-N}$), $\blacklozenge 70\%$ for total N, $\blacklozenge 104\%$ for dissolved P, $\blacklozenge 110\%$ for total P, and $\blacklozenge 53\%$ for total suspended sediments (TSS). Al-Qurashi et al. (2008) defined acceptable performance for flow simulations as being within 30% of observed values for KINEROS/AGWA (Goodrich et al., 2012). For performance measure d , Krause et al. (2005) stated that high values of d (over 0.65) were reported even for poor model fits. In this article, the minimum d value obtained as reported in literature was 0.60, and the overall minimum R^2 value reported in literature and used in the meta-analysis in this article was 0.18. Such low values do not provide much information about model performance and, similar to $NSE < 0.0$, can indicate that the mean observed value is a better predictor than the simulated value.

Thus, in this article, $R^2 < 0.18$, $NSE < 0.0$, $PBIAS = \blacklozenge 30\%$ for flow, $PBIAS = \blacklozenge 55\%$ for sediments, $PBIAS = \blacklozenge 70\%$ for nutrients, and $d < 0.60$ represent unacceptable model performance.

Additional Considerations

The recommendations for model PMs and their corresponding PEC presented in the previous section apply to the typical case of continuous, long-term simulation for the given output responses at specified spatial and temporal scales (table 9). However, because of the diversity of modeling applications, these

recommendations may be adjusted based on the quality and quantity of available measured data, spatial and temporal scales, and project scope and magnitude. It is also important to note that the recommended PMs are based only on the measures reported primarily in the Moriasi et al. (2012) special collection. Therefore, we have provided some additional considerations in this subsection to assist users in their calibration and validation efforts.

The PEC results presented herein are based on a meta-analysis of a selection of published data. As mentioned earlier, this body of data is not all-inclusive; this work can be extended by including data from a more extensive body of literature. However, in order to maintain the integrity of the database, article selection and data collection must be subject to the same considerations and follow the same procedures as outlined in this work. It is also important to note that substantial advances have been made in model calibration and validation such that it is now possible to obtain far better model performance and parameter representation than was possible at its nascence. Thus, we do not recommend the inclusion of historical and early development and application works, as resulting criteria may not be representative of the current state-of-the-art. We suggest using works only from the last 20 years.

A major limitation of the meta-analysis is the exclusion of unpublished data. In further extending the analysis, we recommend, inasmuch as is possible, identification and inclusion of unpublished material that fit all other criteria as outlined under key considerations in the ♦Meta-analysis of Performance Data♦ subsection. The use of only published material in this work has its strengths and weaknesses; while the results are based on data that has undergone a thorough quality assurance and quality review via the peer-review process, a weakness is that typically only good results (with the best performance data values) are published, likely contributing to the lack of distinction among temporal scales. This effect might not be discernible at other levels of analysis since the datasets at those levels are much smaller.

Finally, we recommend presenting summary statistics such as the mean, median, percentiles, and standard deviation of the observed and simulated response outputs. This information is useful and can provide benchmarks for follow up studies.

Residual Analysis

The residual (or error) is the difference between individual observed and simulated values; these values represent the uncertainty of the simulation. Ideally, the residuals should be close to zero and normally distributed. Any skew indicates a systematic bias, which could be potentially resolved by further calibration. Bennett et al. (2013) observed that residual analysis was an important part of model evaluation. They recommended using residual or QQ plots to examine any systematic divergence from zero. Residual plots are graphs of the residuals against time or space, which are useful in identifying any systematic bias. In a QQ plot, quantiles of the residuals are plotted against Gaussian quantiles. This is helpful in determining if the distribution of residuals is normal. Jain and Sudheer (2008) demonstrated that residual analysis, such as checking for homoscedasticity (unsystematic variance), could result in additional insight and improved model evaluations. In addition to graphical analysis, Bennett et al. (2013) recommended calculating the MSE or RMSE of the residuals for a quantitative evaluation.

Despite its documented advantages, residual analysis continues to be a rarely used and/or sparsely reported practice in the modeling literature. Guidelines are needed for simplifying and integrating residual analysis into H/WQ model performance evaluation.

Quality and Quantity of Measured Data

The quality of measured data should be considered in evaluating model calibration and validation performance whenever such information is available (Harmel et al., 2006). According to Harmel et al. (2006), measured data are obtained under best-case, typical, and worst-case data quality scenarios. The best-case scenario represents procedures used with a concentrated effort in quality assurance/quality control (QA/QC),

unconstrained by financial and personnel resource limitations, and in ideal hydrologic conditions. The typical scenario represents procedures conducted with a moderate effort at QA/QC and under typical hydrologic conditions. The worst-case scenario represents data measurements conducted with minimal attention to QA/QC, with limited financial and personnel resources, and in difficult hydrologic conditions. Harmel and Smith (2007) provide modified NSE, d , RMSE, and MAE statistics that account for measurement uncertainty. The recommended model PEC presented herein are for data of typical scenario quality. PEC should be stricter when data of best-case scenario quality are available and more relaxed where uncertainty is high (Moriassi et al., 2007). In such cases, however, users should not over-calibrate their models to obtain values of statistical performance measures better than the uncertainty of the available measured data. Harmel et al. (2010) provide adjustments that can be made to statistical PMs based on uncertainty in measured and simulated data. Alternative measures, such as comparison of means and other graphical PMs such as percentiles and frequency distributions, may be more appropriate for measured datasets derived from either incomplete or low-frequency sampling (Moriassi et al., 2007).

Spatial and Temporal Scale of Study

The recommended PEC are intended for field- to watershed-scale modeling studies and mainly for one or more temporal scales (daily, monthly, and annual) depending on the statistical PMs used and the model output response. More strict PEC are recommended for point to plot scale studies in which there is less complexity of the processes involved and less uncertainty in model inputs (Guzman et al., 2015) due to the small spatial scale (Baffaut et al., 2015). For example, Ma et al. (2012) defined $NSE > 0.70$ and $R^2 > 0.80$ as acceptable model performance values for RZWQM2. It is also necessary to adjust the PEC as the temporal scale changes, utilizing stricter PEC as the evaluation temporal scale decreases from hourly to daily to annual (Moriassi et al., 2007).

Project Scope, Magnitude, and Intended Purpose

Moriassi et al. (2007) discussed the effects of scope and magnitude of the modeling project on model PEC, which should be taken into account when assessing model performance. More stringent PEC are recommended for projects that involve potentially large consequences, while the PEC may be relaxed for proof-of-concept studies. Similarly, Harmel et al. (2014) provided criteria for interpreting model results considering general intended use categories, which include exploratory, planning, and regulatory/legal.

Calibration vs. Validation Performance Criteria

Although prior studies have recommended different PEC for calibration and validation periods (e.g., Moriassi et al., 2007), and our analyses showed significant differences in reported values, this should not be the case. Based on discussions in Sorooshian and Gupta (1995) and Sorooshian (1983), this occurrence in some cases points to inaccuracies in process representation. In other cases, differences in performance during the calibration and validation periods may indicate substantially different climate (Van Liew and Garbrecht, 2003) and land use data (Pai and Saraswat, 2011) and/or the need for further calibration. Thus, the recommended model PEC in this article apply for both the calibration and validation periods. It is also essential to use observed calibration and validation data at spatial and temporal scales that are consistent with the model computations; otherwise, a justification should be provided (Baffaut et al., 2015; Daggupati et al., 2015b).

Framework for Updating Recommended Model Performance Measures and Evaluation Criteria

This initial meta-analysis sets the stage for a more comprehensive meta-analysis including a broader range of articles (including unpublished material) and covering a larger suite of models. To assist with this future endeavor, we present a framework for determining recommended model PMs and their corresponding PEC.

The framework consists of (1) reviewing current modeling literature to determine the PMs used and collect study-specific calibration and validation data as reported and (2) developing PEC for the recommended PMs based on a meta-analysis of a comprehensive dataset collected from published and unpublished sources while taking into account all key considerations described herein. The scope and limitations of the recommended PEC in this article have been clearly defined in prior sections but can be updated as more information becomes available. For future work, we recommend using performance data values reported for other models, for different output responses, and at various spatial and temporal scales both from published and unpublished literature. In addition, reported study-specific graphical PMs need to be recorded and discussed in depth.

We have established a database with an inventory of reported model performance values and respective study details (e.g., spatial scales, outputs, objective functions) to enable modelers to query and develop custom model PEC better suited to their study goals. This database can be extended frequently as H/WQ model PMs and related PEC continue to evolve and when new understandings of modeling science arise. We intend to make this database available in an open and user-friendly format to provide opportunities for updates through crowd-sourcing. The analysis framework and the developed database will enable modifications of the recommended PMs and PEC as more information is obtained.

Demonstration of Recommended Model Performance Measures and Criteria

An example case study was conducted with a hypothetical watershed-scale H/WQ model. The model was calibrated at the outlet for stream flow on a daily temporal scale for ten years (2001 to 2010). The model name and the study location are not mentioned here to emphasize the generic nature of the guidelines. Figure 4 and table 10 show the graphical and statistical performance of the model based on the recommended PMs.

Since this is a daily temporal scale, ten-year evaluation, the recommended graphical PMs are the scatter plot and flow duration curves (fig. 4). Note that a time series graph was not recommended in this case because of the large dataset. The slope and intercept values are provided on the scatter plot based on the least square regression line. The slope of the line is close to a value of one, while the intercept is close to a value of zero, indicating good model performance. The flow duration curve shows that model predictions were close to the observed data for all flow regimes, although the model tended to underestimate the observed data during low flows (>80% probability), slightly overestimate during medium flows (>20% and <50% probability), and had a good agreement during high flows (>10%). By using this figure, a modeler and end user can easily visualize model performance and further identify parameters that can be tweaked to improve performance. For instance, in this case, parameters related to base flow can be adjusted, allowing the model to simulate slightly higher low flows.

Based on the statistical PMs, we can say that the model adequately captured the mean and standard deviation of the daily flow rates. Using the performance values in table 9, we can say that model performance was ♦satisfactory♦ based on R^2 and NSE, ♦not satisfactory♦ based on the PBIAS of -16%, and satisfactory based on the RSR of 0.63 (Moriassi et al., 2007). Adjustments can be made to model parameters to obtain better agreement among the PMs.

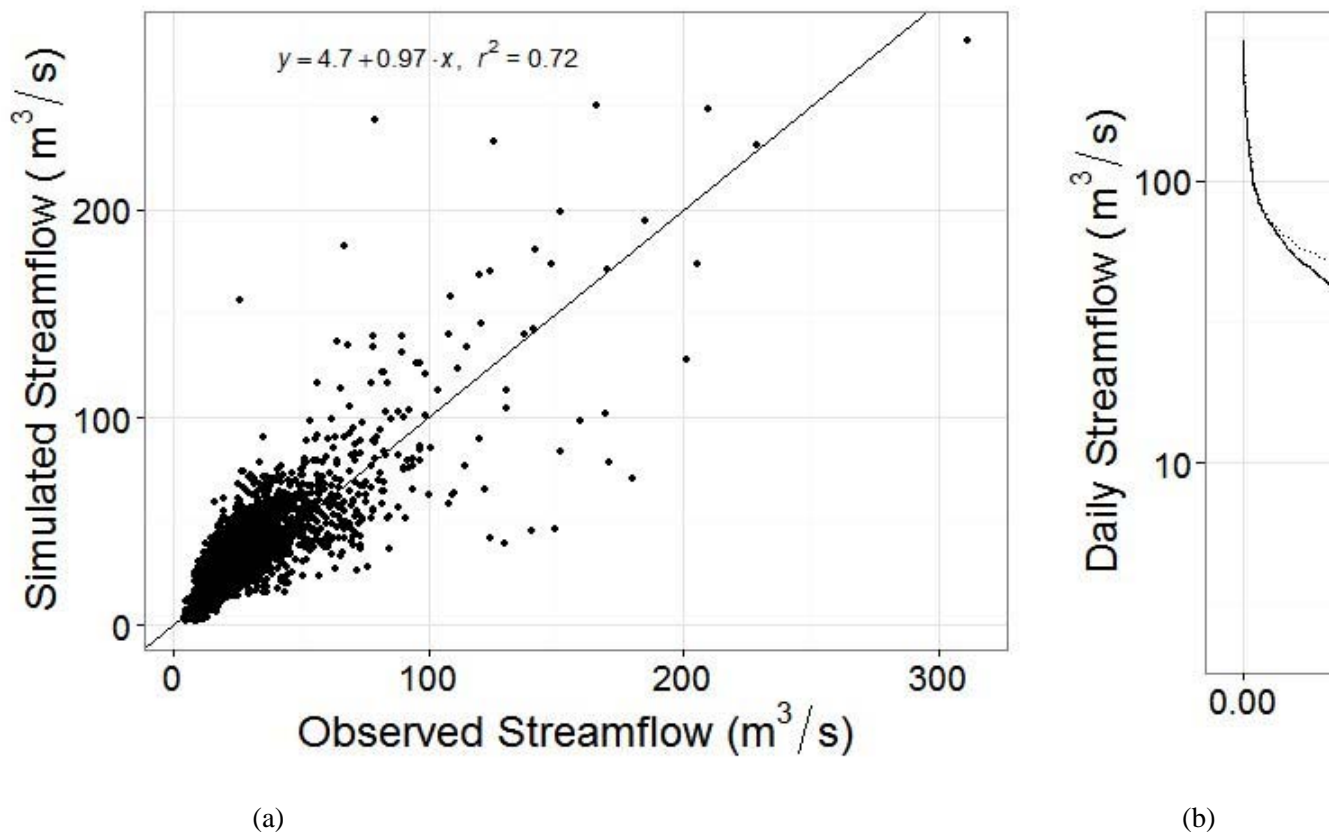


Figure 4. Graphical performance measures of a hypothetical model: (a) scatter plot and (b) flow duration curve.

Table 10. Statistical performance evaluation criteria of a hypothetical model.									
Average		Standard Deviation		Statistics					
Measured	Simulated	Measured	Simulated	R^2	PBIAS (%)	NSE	RSR	RMSE	
24.4	28.3	21.0	23.9	0.72 (slope 0.97, intercept 4.7) (Satisfactory)	-16 (Not satisfactory)	0.60 (Satisfactory)	0.63 (Satisfactory; Moriassi et al., 2007)		13.2

Although H/WQ models provide outputs in various file formats, performance evaluation is typically performed using a spreadsheet. However, setting up a spreadsheet to calculate the numerous graphical and statistical PMs can be a tedious task and prone to errors. Therefore, to support the task of model performance evaluation, a Microsoft Excel spreadsheet was developed (available at http://bit.ly/NRES_SW10715). The objectives of the spreadsheet are to (1) demonstrate the various statistical and graphical PMs discussed in the case study and (2) provide a starting point for H/WQ model users to conduct model performance evaluation.

In situations with conflicting performance ratings, we recommend that those differences be clearly described. For example, if simulation for one output variable in one watershed produces unbalanced performance ratings of \blacklozenge satisfactory \blacklozenge for R^2 and \blacklozenge unsatisfactory \blacklozenge for d for field-scale flow simulation, then the overall performance should be described conservatively as \blacklozenge unsatisfactory \blacklozenge for that one study area and that one model response output. However, we recommend that users describe model performance with respect to the degree of collinearity between simulated and measured data (R^2) as \blacklozenge satisfactory \blacklozenge and with respect to prediction error (d) as \blacklozenge unsatisfactory \blacklozenge . Similarly, if performance ratings differ for various field- and watershed-scale studies and/or response outputs, then those differences need to be clearly described.

Summary and Conclusions

This is one of nine topic-specific articles in a special collection whose main goal is to provide recommendations that, together with recommendations by Harmel et al. (2014), will contribute toward the development of ASABE engineering practices for calibration and validation of H/WQ models. In this research, articles in the Moriasi et al. (2012) special collection were synthesized with respect to performance measures (PMs) and performance evaluation criteria (PEC). In addition, a detailed literature review centered on graphical and statistical PMs used by models described in the special collection was carried out to determine PMs to recommend for use. Further, an initial meta-analysis of performance data reported in literature (outside of the special collection) was performed to establish PEC for various PMs. Data were collected from articles published from 1992 to 2013; 93% were published in and after 2000, and 53% were published after 2007. Finally, specific guidelines for model performance evaluation were established based on the synthesis and results of the meta-analysis. Additional considerations were also presented to allow users to adjust recommended PMs and/or associated PEC to their specific needs. A framework for determining recommended model PMs and their corresponding PEC, based on a more comprehensive meta-analysis, was presented.

Based on the synthesis, we recommend that a combination of multiple graphical and statistical PMs be used for evaluating model performance. Recommended graphical PMs include time series, scatter plots, cumulative distribution, flow and load duration, and maps, while the recommended statistical PMs include R^2 (in conjunction with slope and intercept of the pertinent regression line), NSE, d , RMSE (together with RSR), and PBIAS.

In this study, we do not go further into specifying PEC based on watershed size, although further work would be needed in this regard. However, the results strongly suggest the need to provide PEC at different scales; therefore, we provide separate PEC for the watershed scale and the field scale. We do not provide (or even recommend) separate PEC for calibration and validation periods. Based on the meta-analysis results, previous PEC reported in the literature, and our modeling experience, recommended PEC are presented in table 9. In general, model performance can be judged \blacklozenge satisfactory \blacklozenge for flow simulations if monthly $R^2 > 0.70$ and $d > 0.75$ for field-scale models and daily, monthly, or annual $R^2 > 0.60$, $NSE > 0.50$, and $PBIAS = \blacklozenge 15\%$ for watershed-scale models. Additionally, model performance can be judged \blacklozenge satisfactory \blacklozenge if monthly $R^2 > 0.40$ and $NSE > 0.45$ and daily, monthly, or annual $PBIAS = \blacklozenge 20\%$ for sediment; monthly $R^2 > 0.40$ and $NSE > 0.35$ and daily, monthly, or annual $PBIAS = \blacklozenge 30\%$ for P; and monthly $R^2 > 0.30$ and $NSE > 0.35$ and daily, monthly, or annual $PBIAS = \blacklozenge 30\%$ for N. For RSR, we recommend that the PEC proposed by Moriasi et al. (2007) be used until new PEC are developed. These PEC, which apply to calibration and validation periods, may be adjusted to be more or less strict based on considerations of the quality and quantity of available measured data, spatial and temporal scales, and project scope, magnitude, and intended purpose. As more data become available and as new PMs are developed and used more frequently, the recommended PMs and their corresponding general PEC can be adjusted based on the framework developed in this study.

However, we consider some values of the recommended statistical PMs to be unacceptable beyond certain reasonable ranges. Thus, in this article, $R^2 < 0.18$, $NSE < 0.0$, $PBIAS = \blacklozenge 30\%$ for flow, $PBIAS = \blacklozenge 55\%$ for sediment, $PBIAS = \blacklozenge 70\%$ for nutrients, and $d < 0.60$ represent unacceptable model performance. An example case study and an Excel spreadsheet are provided to illustrate the application of the recommended PMs and the corresponding developed PEC guidelines.

The guidelines developed in this study go beyond the scope of those provided by Moriasi et al. (2007), which were limited to NSE, PBIAS (Gupta et al., 1999), and RSR for stream flow, sediment, and nutrient (N and P) simulations at a monthly temporal scale and watershed spatial scale. In this study, PEC for R^2 were added and PEC for NSE were disaggregated by output parameter (flow, sediment, N/P), and limits were adjusted based on current data. Limits were also adjusted for PBIAS for each output parameter, and some PEC were

explicitly extended to daily and annual scales. In addition, PEC for R^2 and d were added for ADAPT. These current results provide updated guidance on performance measures and corresponding performance evaluation criteria for calibrating and validating hydrologic and water quality models.

Acknowledgements

The authors are grateful to Dr. Dharmendra Saraswat, Dr. Colleen Rossi, Dr. Sanjay Shukla, and Dr. Prasanna Gowda for their initial support of these efforts. We also thank all those who reviewed the manuscript.

References

- Al-Qurashi, A., McIntyre, N., Wheeler, H., & Unkrich, C. (2008). Application of the Kinos2 rainfall-runoff model to an arid catchment in Oman. *J. Hydrol.*, 355(1), 91-105. <http://dx.doi.org/10.1016/j.jhydrol.2008.03.022>.
- Arnold, J., Moriasi, D., Gassman, P., Abbaspour, K., White, M., Srinivasan, R., Santhi, C., Harmel, R. D., van Griensven, A., Van Liew, M. W., Kannan, N., & Jha, M. K. (2012). SWAT: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1491-1508. <http://dx.doi.org/10.13031/2013.42256>.
- Arnold, J. G., Youssef, M. A., Yen, H., White, M. J., Sheshukov, A. Y., Sadeghi, A. M., Moriasi, D. N., Steiner, J. L., Amatya, D. M., Skaggs, R. W., Haney, E. B., Jeong, J., Arabi, M., & Gowda, P. H. (2015). Hydrological processes and model representation: Impact of soft data on calibration. *Trans. ASABE*, 58(6), 1637-1660. <http://dx.doi.org/10.13031/trans.58.10726>.
- ASCE. (1993). Criteria for evaluation of watershed models. *J. Irrig. Drain. Eng.*, 119(3), 429-442. [http://dx.doi.org/10.1061/\(ASCE\)0733-9437\(1993\)119:3\(429\)](http://dx.doi.org/10.1061/(ASCE)0733-9437(1993)119:3(429)).
- Baffaut, C., Dabney, S. M., Smolen, M. D., Youssef, M. A., Bonta, J. V., Chu, M. L., Guzman, J. A., Shedekar, V., Jha, M. K., & Arnold, J. G. (2015). Hydrologic and water quality modeling: Spatial and temporal considerations. *Trans. ASABE*, 58(6), 1661-1680. <http://dx.doi.org/10.13031/trans.58.10714>.
- Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D., & Andreassian, V. (2013). Characterising performance of environmental models. *Environ. Model. Software*, 40, 1-20. <http://dx.doi.org/10.1016/j.envsoft.2012.09.011>.
- Biondi, D., Freni, G., Iacobellis, V., Mascaro, G., & Montanari, A. (2012). Validation of hydrological models: Conceptual basis, methodological approaches, and a proposal for a code of practice. *Phys. Chem. Earth*, 42-44, 70-76. <http://dx.doi.org/10.1016/j.pce.2011.07.037>.
- Black, D. C., Wallbrink, P. J., & Jordan, P.W. (2014). Towards best practice implementation and application of models for analysis of water resources management scenarios. *Environ. Model. Software*, 52, 136-148. <http://dx.doi.org/10.1016/j.envsoft.2013.10.023>.
- Bland, M. (2000). *An Introduction to Medical Statistics*. Oxford, U.K.: Oxford University Press.
- Beven, K., & Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *J. Hydrol.*, 249(1), 11-29. [http://dx.doi.org/10.1016/S0022-1694\(01\)00421-8](http://dx.doi.org/10.1016/S0022-1694(01)00421-8).
- Bottcher, A. D. B., Whiteley, B. J., James, A. I., & Hiscock, J. G. (2012). Watershed Assessment Model (WAM): Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1367-1383. <http://dx.doi.org/10.13031/2013.42248>.

Brown, G. W., & Mood, A. M. (1951). On median tests for linear hypotheses. In *Proc. 2nd Berkeley Symp.: Mathematical Statistics and Probability* (pp. 159-166). Berkeley, Cal.: University of California Press.

Cho, H., & Olivera, F. (2009). Effect of the spatial variability of land use, soil type, and precipitation on streamflows in small watersheds. *JAWRA*, 45(3), 1423-1431. <http://dx.doi.org/10.1111/j.1752-1688.2009.00315.x>.

Coffey, M., Workman, S., Taraba, J., & Fogle, A. (2004). Statistical procedures for evaluating daily and monthly hydrologic model predictions. *Trans. ASAE*, 47(1), 59-68. <http://dx.doi.org/10.13031/2013.15870>.

Daggupati, P., Douglas-Mankin, K.R., Sheshukov, A. Y., Barnes, P. L., & D. L. Devlin. (2011). Field-level targeting using SWAT: Mapping output from HRUs to fields and assessing limitations of GIS input data, *Trans. ASABE*, 54(2), 501-514. <http://dx.doi.org/10.13031/2013.36453>.

Daggupati, P., Sheshukov, A. Y. & Douglas-Mankin, K. R. (2014). Evaluating ephemeral gullies with a process-based topographic index model. *Catena*, 113, 177-186. <http://dx.doi.org/10.1016/j.catena.2013.10.005>.

Daggupati, P., Yen, H., White, M. J., Srinivasan, R., Arnold, J. G., Keitzer, C. S., & Sowa, S. P. (2015a). Impact of model development decision on hydrological processes and streamflow. *Hydrol. Proc.* 29(26), 5307-5320. <http://dx.doi.org/10.1002/hyp.10536>.

Daggupati, P., Pai, N., Ale, S., Douglas-Mankin, K. R., Zeckoski, R. W., Jeong, J., Parajuli, P. B., Saraswat, D., & Youssef, M. A. (2015). A recommended calibration and validation strategy for hydrologic and water quality models. *Trans. ASABE*, 58(6), 1705-1719. <http://dx.doi.org/10.13031/trans.58.10712>.

Diekkrger, B., Sndgerath, D., Kersebaum, K., & McVoy, C. (1995). Validity of agroecosystem models: A comparison of results of different models applied to the same data set. *Ecol. Model.*, 81(1-3), 3-29. [http://dx.doi.org/10.1016/0304-3800\(94\)00157-D](http://dx.doi.org/10.1016/0304-3800(94)00157-D).

Doherty, J., & Johnston, J. M. (2003). Methodologies for calibration and predictive analysis of a watershed model. *JAWRA*, 39(2), 251-265. <http://dx.doi.org/10.1111/j.1752-1688.2003.tb04381.x>.

Douglas-Mankin, K., Srinivasan, R., & Arnold, J. (2010). Soil and Water Assessment Tool (SWAT) model: Current developments and applications. *Trans. ASABE*, 53(5), 1423-1431. <http://dx.doi.org/10.13031/2013.34915>.

Douglas-Mankin, K. R., Daggupati, P., Sheshukov, A. Y., & Barnes, P. L. (2013). Paying for sediment: Field-scale conservation practice targeting, funding, and assessment using SWAT. *J. Soil Water Cons.*, 68(1), 41-51. <http://dx.doi.org/10.2489/jswc.68.1.41>.

Duda, P. B., Hummel Jr., P. R., Donigian Jr., A. S., & Imhoff, J. C. (2012). BASINS/HSPF: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1523-1547. <http://dx.doi.org/10.13031/2013.42261>.

Eckhardt, K., & Arnold, J. (2001). Automatic calibration of a distributed catchment model. *J. Hydrol.*, 251(1), 103-109. [http://dx.doi.org/10.1016/S0022-1694\(01\)00429-2](http://dx.doi.org/10.1016/S0022-1694(01)00429-2).

Egger, M., & Smith, G. D. (1997). Meta-analysis: Potentials and promise. *British Med. J.*, 315(7119), 1371-1374. <http://dx.doi.org/10.1136/bmj.315.7119.1371>.

Engel, B., Storm, D., White, M., Arnold, J., & Arabi, M. (2007). A hydrologic/water quality model application protocol. *JAWRA*, 43(5), 1223-1236. <http://dx.doi.org/10.1111/j.1752-1688.2007.00105.x>.

Essaid, H. I., Zamora, C. M., McCarthy, K. A., Vogel, J. R., & Wilson, J. T. (2008). Using heat to characterize streambed water flux variability in four stream reaches. *J. Environ. Qual.*, 37(3), 1010-1023. <http://dx.doi.org/10.2134/jeq2006.0448>.

- Fernandez, G. P., Chescheir, G. M., Skaggs, R. W., & Amatya, D. M. (2005). Development and testing of watershed-scale models for poorly drained soils. *Trans. ASAE*, 48(2), 639-652. <http://dx.doi.org/10.13031/2013.18323>.
- Finsterle, S., Kowalsky, M. B., & Pruess, K. (2012). TOUGH: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1275-1290. <http://dx.doi.org/10.13031/2013.42240>.
- Flanagan, D. C., Frankenberger, J. R., & Ascoug II, J. C. (2012). WEPP: Model use, calibration and validation. *Trans. ASABE*, 55(4), 1463-1477. <http://dx.doi.org/10.13031/2013.42254>.
- Flerchinger, G. N., Caldwell, T. G., Cho, J., & Hardegree, S. (2012). Simultaneous Heat and Water (SHAW): Model use, calibration, and validation. *Trans. ASABE* 55(4), 1395-1411. <http://dx.doi.org/10.13031/2013.42250>.
- Gan, T. Y., & Biftu, G. F. (1996). Automatic calibration of conceptual rainfall-runoff models: Optimization algorithms, catchment conditions, and model structure. *Water Resources Res.*, 32(12), 3513-3524. <http://dx.doi.org/10.1029/95WR02195>.
- Gan, T. Y., Dlamini, E. M., & Biftu, G. F. (1997). Effects of model complexity and structure, data quality, and objective functions on hydrologic modeling. *J. Hydrol.*, 192(1), 81-103. [http://dx.doi.org/10.1016/S0022-1694\(96\)03114-9](http://dx.doi.org/10.1016/S0022-1694(96)03114-9).
- Gassman, P. W., Reyes, M. R., Green, C. H., & Arnold, J. G. (2007). The soil and water assessment tool: Historical development, applications, and future directions. *Trans. ASABE* 50(4), 1211-1250. <http://dx.doi.org/10.13031/2013.23637>.
- Gitau, M. W., Gburek, W. J. & Jarrett, A. R. (2005). A tool for estimating best management practice effectiveness for phosphorus pollution control. *J. Soil Water Cons.*, 60(1), 1-10.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educ. Res.*, 5(10), 3-8. <http://dx.doi.org/10.3102/0013189X005010003>.
- Goodrich, D. C., Burns, I. S., Unkrich, C. L., Semmens, D. J., Guertin, D. P., Hernandez, M., Yatheendradas, S., Kennedy, J. R., & Levick, L. R. (2012). KINEROS2/AGWA: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1561-1574. <http://dx.doi.org/10.13031/2013.42264>.
- Gowda, P. H., Mulla, D. J., Desmond, E. D., Ward, A. D., & Moriasi, D. N. (2012). ADAPT: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1345-1352. <http://dx.doi.org/10.13031/2013.42246>.
- Gupta, H. V., Sorooshian, S., & Yapo, P. O. (1998). Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Res.*, 34(4), 751-763. <http://dx.doi.org/10.1029/97WR03495>.
- Gupta, H. V., Sorooshian, S., & Yapo, P. O. (1999). Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. *J. Hydrol. Eng.*, 4(2), 135-143. [http://dx.doi.org/10.1061/\(ASCE\)1084-0699\(1999\)4:2\(135\)](http://dx.doi.org/10.1061/(ASCE)1084-0699(1999)4:2(135)).
- Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.*, 377(1), 80-91. <http://dx.doi.org/10.1016/j.jhydrol.2009.08.003>.
- Guzman, J. A., Shirmohammadi, A., Sadeghi, A. M., Wang, X., Chu, M. L., Jha, M. K., Parajuli, P. B., Harmel, R. D., Khare, Y., & Hernandez, J. (2015). Uncertainty considerations in calibration and validation of hydrologic and water quality models. *Trans. ASABE*, 58(6), 1745-1762. <http://dx.doi.org/10.13031/trans.58.10710>.

Hansen, S., Abrahamsen, P., Petersen, C. T., & Styczen, M. (2012). Daisy: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1317-1335. <http://dx.doi.org/10.13031/2013.42244>.

Harmel, D. R., & Smith, P. K. (2007). Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling. *J. Hydrol.*, 337(3), 326-336. <http://dx.doi.org/10.1016/j.jhydrol.2007.01.043>.

Harmel, R. D., Cooper, R. J., Slade, R. M., Haney, R. L., & Arnold, J. G. (2006). Cumulative uncertainty in measured stream flow and water quality data for small watersheds. *Trans. ASABE*, 49(3), 689-701. <http://dx.doi.org/10.13031/2013.20488>.

Harmel, R. D., Smith, P. K., & Migliaccio, K. W. (2010). Modifying goodness-of-fit indicators to incorporate both measurement and model uncertainty in model calibration and validation. *Trans. ASABE* 53(1), 55-63. <http://dx.doi.org/10.13031/2013.29502>.

Harmel, R. D., Smith, P. K., Migliaccio, K. W., Chaubey, I., Douglas-Mankin, K. R., Benham, B., Shukla, S., Muoz-Carpena, R., & Robson, B. J. (2014). Evaluating, interpreting, and communicating performance of hydrologic/water quality models considering intended use: A review and recommendations. *Environ. Model. Software*, 57, 40-51. <http://dx.doi.org/10.1016/j.envsoft.2014.02.013>.

Healy, R. W., & Essaid, H. I. (2012). VS2DI: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1249-1260. <http://dx.doi.org/10.13031/2013.42238>.

Hendricks, G., Shukla, S., Martinez, C., & Kiker, G. (2013). Modified model for simulating hydrologic processes for plastic mulch production systems. *J. Irrig. Drain. Eng.*, 139(9), 738-746. [http://dx.doi.org/10.1061/\(ASCE\)IR.1943-4774.0000615](http://dx.doi.org/10.1061/(ASCE)IR.1943-4774.0000615).

Herr, J. W., & Chen, C. W. (2012). WARMF: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1385-1394. <http://dx.doi.org/10.13031/2013.42249>.

Huisman, J., Hubbard, S., Redman, J., & Annan, A. (2003). Measuring soil water content with ground-penetrating radar. *Vadose Zone J.*, 2(4), 476-491. <http://dx.doi.org/10.2136/vzj2003.4760>.

Hunt, M. 1997. *How Science Takes Stock: The Story of Meta-Analysis*. New York, N.Y.: Russell Sage Foundation.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-Analysis: Cumulating Research Findings across Studies*. Beverly Hills, Cal.: Sage Publications.

Huth, N. I., Bristow, K. L., & Verburg, K. (2012). SWIM3: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1303-1313. <http://dx.doi.org/10.13031/2013.42243>.

Jaber, F. H., & Shukla, S. (2012). MIKE SHE: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1479-1489. <http://dx.doi.org/10.13031/2013.42255>.

Jaber, F. H., Shukla, S., & Srivastava, S. (2006). Recharge, upflux, and water table response for shallow water table conditions. *Hydrol. Proc.*, 20(9), 1895-1907. <http://dx.doi.org/10.1002/hyp.5951>.

Jain, S. K., & Sudheer, K. P. (2008). Fitting of hydrologic models: A close look at the Nash-Sutcliffe index. *J. Hydrol. Eng.*, 13(10), 981-986. [http://dx.doi.org/10.1061/\(ASCE\)1084-0699\(2008\)13:10\(981\)](http://dx.doi.org/10.1061/(ASCE)1084-0699(2008)13:10(981)).

Jansson, P. (2012). COUP Model: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1337-1346. <http://dx.doi.org/10.13031/2013.42245>.

Jarvis, N., & Larsbo, M. (2012). MACRO (v5.2): Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1413-1423. <http://dx.doi.org/10.13031/2013.42251>.

- Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Res.*, 42(3), W03S04. <http://dx.doi.org/10.1029/2005WR004362>.
- Knisel, W. G., & Douglas-Mankin, K. R. (2012). CREAMS/GLEAMS: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1291-1302. <http://dx.doi.org/10.13031/2013.42241>.
- Krause, P., Boyle, D., & Bese, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Adv. Geosci.*, 5, 89-97. <http://dx.doi.org/10.5194/adgeo-5-89-2005>.
- Krabel, R., Sun, Q., Ingwersen, J., Chen, X., Zhang, F., Moller, T., & Rmheld, V. (2010). Modelling water dynamics with DNDC and DAISY in a soil of the North China Plain: A comparative study. *Environ. Model. Software*, 25(4), 583-601. <http://dx.doi.org/10.1016/j.envsoft.2009.09.003>.
- Legates, D. R., & McCabe, G. J. (1999). Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation. *Water Resources Res.*, 35(1), 233-241. <http://dx.doi.org/10.1029/1998WR900018>.
- Light, R. J., & Pillemer, D. B. (1984). *Summing Up: The Science of Reviewing Research*. Cambridge, Mass.: Harvard University Press.
- Light, R., & Smith, P. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educ. Rev.*, 41(4), 429-471. <http://dx.doi.org/10.17763/haer.41.4.437714870334w144>.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-Analysis*. Thousand Oaks, Cal.: Sage Publications.
- Loague, K., & Green, R. E. (1991). Statistical and graphical methods for evaluating solute transport models: Overview and application. *J. Contam. Hydrol.*, 7(1), 51-73. [http://dx.doi.org/10.1016/0169-7722\(91\)90038-3](http://dx.doi.org/10.1016/0169-7722(91)90038-3).
- Lyons, L. C. (1998). Meta-analysis: Methods of accumulating results across research domains. Retrieved from www.lyonsmorris.com/MetaA/macalc/MApaper.pdf.
- Ma, L., Ahuja, L., Nolan, B., Malone, R., Trout, T., & Qi, Z. (2012). Root Zone Water Quality Model (RZWQM 2): Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1425-1446. <http://dx.doi.org/10.13031/2013.42252>.
- Malone, R. W., Yagow, G., Baffaut, C., Gitau, M. W., Qi, Z., Amatya, D. M., Parajuli, P. B., Bonta, J. V., & Green, T. R. (2015). Parameterization guidelines and considerations for hydrologic models. *Trans. ASABE*, 58(6), 1681-1703. <http://dx.doi.org/10.13031/trans.58.10709>.
- McCuen, R. H., Knight, Z., & Cutter, A. G. (2006). Evaluation of the Nash-Sutcliffe efficiency index. *J. Hydrol. Eng.*, 11(6), 597-602. [http://dx.doi.org/10.1061/\(ASCE\)1084-0699\(2006\)11:6\(597\)](http://dx.doi.org/10.1061/(ASCE)1084-0699(2006)11:6(597)).
- Moriasi, D., Arnold, J., Van Liew, M., Bingner, R., Harmel, R., & Veith, T. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE*, 50(3), 885-900. <http://dx.doi.org/10.13031/2013.23153>.
- Moriasi, D., Wilson, B., Douglas-Mankin, K., Arnold, J., & Gowda, P. (2012). Hydrologic and water quality models: Use, calibration, and validation. *Trans. ASABE*, 55(4), 1241-1247. <http://dx.doi.org/10.13031/2013.42265>.
- Mutiti, S., & Levy, J. (2010). Using temperature modeling to investigate the temporal variability of riverbed hydraulic conductivity during storm events. *J. Hydrol.*, 388(3), 321-334. <http://dx.doi.org/10.1016/j.jhydrol.2010.05.011>.

- Nash, J., & Sutcliffe, J. (1970). River flow forecasting through conceptual models: Part I. A discussion of principles. *J. Hydrol.*, 10(3), 282-290. [http://dx.doi.org/10.1016/0022-1694\(70\)90255-6](http://dx.doi.org/10.1016/0022-1694(70)90255-6).
- Pai, N., & D. Saraswat. (2011). SWAT2009_LUC: A tool to activate the land use change module in SWAT 2009. *Trans. ASABE*, 54(5), 1649-1658.
- Pai, N., Saraswat, D., & Daniels, M. (2011). Identifying priority subwatersheds in the Illinois river drainage area in Arkansas watershed using a distributed modeling approach. *Trans. ASABE*, 54(6), 2181-2196. <http://dx.doi.org/10.13031/2013.40657>.
- Palosuo, T., Kersebaum, K. C., Angulo, C., Hlavinka, P., Moriondo, M., Olesen, J. E., Patil, R., Ruget, F., Rumbaur, C., Taká, J., Trnka, M., Bindi, M., Caldag, B., Ewert, F., Ferrise, R., Mirschel, W., Saylan, L., Siska, B., Rötter, R. (2011). Simulation of winter wheat yield and its variability in different climates of Europe: A comparison of eight crop growth models. *European J. Agron.*, 35(3), 103-114. <http://dx.doi.org/10.1016/j.eja.2011.05.001>.
- Parajuli, P. B., Nelson, N. O., Frees, L. D., & Mankin, K. R. (2009). Comparison of AnnAGNPS and SWAT model simulation results in USDA-CEAP agricultural watersheds in south-central Kansas. *Hydrol. Proc.*, 23(5), 748-763. <http://dx.doi.org/10.1002/hyp.7174>.
- Pfannerstill, M., Guse, B., & Fohrer, N. (2014). Smart low flow signature metrics for an improved overall performance evaluation of hydrological models. *J. Hydrol.*, 510, 447-458. <http://dx.doi.org/10.1016/j.jhydrol.2013.12.044>.
- Popov, E. G. (1979). *Gidrologicheskie Prognozy (Hydrological Forecasts)*. Leningrad, Russia. As cited in Van Liew, M. W., & Garbrecht, J. 2003. Hydrologic simulation of the Little Washita river experimental watershed using SWAT. *JAWRA*, 39(2):413-426. <http://dx.doi.org/10.1111/j.1752-1688.2003.tb04395.x>.
- Pushpalatha, R., Perrin, C., Le Moine, N., & Andréassian, V. (2012). A review of efficiency criteria suitable for evaluating low-flow simulations. *J. Hydrol.*, 420-421, 171-182. <http://dx.doi.org/10.1016/j.jhydrol.2011.11.055>.
- Ramanarayanan, T., Williams, J., Dugas, W., Hauck, L., & McFarland, A. (1997). Using APEX to identify alternative practices for animal waste management. ASAE Paper No. 972209. St. Joseph, Mich.: ASAE.
- Refsgaard, J. C. (1997). Parameterisation, calibration, and validation of distributed hydrological models. *J. Hydrol.*, 198(1-4), 69-97. [http://dx.doi.org/10.1016/S0022-1694\(96\)03329-X](http://dx.doi.org/10.1016/S0022-1694(96)03329-X).
- Reungsang, P., Kanwar, R., & Srisuk, K. (2010). Application of SWAT model in simulating stream flow for the Chi River subbasin II in northeast Thailand. *Trends Res. Sci. Tech.*, 2(1), 23-28.
- Ritter, A., & Muñoz-Carpena, R. (2013). Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *J. Hydrol.*, 480, 33-45. <http://dx.doi.org/10.1016/j.jhydrol.2012.12.004>.
- Santhi, C., Arnold, J. G., Williams, J. R., Dugas, W. A., Srinivasan, R., & Hauck, L. M. (2001). Validation of the SWAT model on a large river basin with point and nonpoint sources. *JAWRA*, 37(5), 1169-1188. <http://dx.doi.org/10.1111/j.1752-1688.2001.tb03630.x>.
- Saraswat, D., Frankenberg, J. R., Pai, N., Ale, S., Daggupati, P., Douglas-Mankin, K. R., & Youssef, M. A. (2015). Hydrologic and water quality models: Documentation and reporting procedures for calibration, validation, and use. *Trans. ASABE*, 58(6), 1787-1797. <http://dx.doi.org/10.13031/trans.58.10707>.
- SAS. 2007. *JMP Statistics and Graphics Guide*. Cary, N.C.: SAS Institute, Inc.

SAS. 2008. JMP Version 8. Cary, N.C.: SAS Institute, Inc.

Sheskin, 2003. *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, Fla.: CRC Press.

Shirmohammadi, A., Chaubey, I., Harmel, R. D., Bosch, D. D., Muñoz-Carpena, R., Dharmasri, C., Sexton, A., Arabi, M., Wolfe, M. L., Frankenberger, J., Graff, C., & Sohrabi, T. M. (2006), Uncertainty in TMDL models. *Trans. ASABE*, 49(4), 1033-1049. <http://dx.doi.org/10.13031/2013.21741>.

Šimuněk, J., van Genuchten, M. T., & Šejna, M. (2012). HYDRUS: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1261-1274. <http://dx.doi.org/10.13031/2013.42239>.

Singh, J., Knapp, H. V., & Demissie, M. (2004). Hydrologic modeling of the Iroquois River watershed using HSPF and SWAT. ISWS CR 2004-08. Champaign, Ill.: Illinois State Water Survey.

Skaggs, R., Youssef, M., & Chescheir, G. (2012). DRAINMOD: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1509-1522. <http://dx.doi.org/10.13031/2013.42259>.

Sorooshian, S. (1983). Surface water hydrology: On-line estimation. *Rev. Geophys.*, 21(3), 706-721. <http://dx.doi.org/10.1029/RG021i003p00706>.

Sorooshian, S., & Gupta, V. K. (1995). Chapter 2: Model calibration. In V. P. Singh (Ed.), *Computer Models of Watershed Hydrology*. Highlands Ranch, Colo.: Water Resources Publications.

Srinivasan, R., Zhang, X., & Arnold, J. (2010). SWAT ungauged: Hydrological budget and crop yield predictions in the upper Mississippi River basin. *Trans. ASABE*, 53(5), 1533-1546. <http://dx.doi.org/10.13031/2013.34903>.

Tuppad, P., Douglas-Mankin, K., Lee, T., Srinivasan, R., & Arnold, J. (2011). Soil and Water Assessment Tool (SWAT) hydrologic/water quality model: Extended capability and wider adoption. *Trans. ASABE*, 54(5), 1677-1684. <http://dx.doi.org/10.13031/2013.39856>.

USEPA. (2002). Guidance for quality assurance project plans for modeling. EPA QA/G-5M Report EPA/240/R-02/007. Washington, D.C.: U.S. Environmental Protection Agency. Retrieved from www.epa.gov/quality/guidance-quality-assurance-project-plans-modeling-epa-qag-5m.

USEPA. (2009). Guidance on the development, evaluation, and application of environmental models. EPA/100/K-09/003. Washington, D.C.: U.S. Environmental Protection Agency. Retrieved from www.epa.gov/crem/library/cred_guidance_0309.pdf.

USEPA. (2010). Economic analysis of final water quality standards for nutrients for lakes and flowing waters in Florida. Washington, D.C.: U.S. Environmental Protection Agency. Retrieved from http://water.epa.gov/lawsregs/rulesregs/upload/florida_econ.pdf.

van der Keur, P., Hansen, S., Schelde, K., & Thomsen, A. (2001). Modification of DAISY SVAT model for potential use of remotely sensed data. *Agric. Forest Meteorol.*, 106(3), 215-231.

van Genuchten, M. T., Šimuněk, J., Leij, F. J., Toride, N., & Šejna, M. (2012). STANMOD: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1355-1368. <http://dx.doi.org/10.13031/2013.42247>.

van Griensven, A., & Bauwens, W. (2003). Multiobjective autocalibration for semidistributed water quality models. *Water Resources Res.*, 39(12), 1348. <http://dx.doi.org/10.1029/2003WR002284>.

Van Liew, M. W., & Garbrecht, J. (2003). Hydrologic simulation of the Little Washita River experimental watershed using SWAT. *JAWRA*, 39(2), 413-426. <http://dx.doi.org/10.1111/j.1752-1688.2003.tb04395.x>.

- Van Liew, M. W., Veith, T. L., Bosch, D. D., & Arnold, J. G. (2007). Suitability of SWAT for the conservation effects assessment project: Comparison on USDA agricultural research service watersheds. *J. Hydrol. Eng.*, 12(2), 173-189. [http://dx.doi.org/10.1061/\(ASCE\)1084-0699\(2007\)12:2\(173\)](http://dx.doi.org/10.1061/(ASCE)1084-0699(2007)12:2(173)).
- Vazquez-Amabile, G., & Engel, B. (2005). Use of SWAT to compute groundwater table depth and streamflow in the Muscatatuck River watershed. *Trans. ASAE*, 48(3), 991-1003. <http://dx.doi.org/10.13031/2013.18511>.
- Wang, X., Williams, J., Gassman, P., Baffaut, C., Izaurrealde, R., Jeong, J., & Kiniry, J. (2012). EPIC and APEX: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1447-1462. <http://dx.doi.org/10.13031/2013.42253>.
- Westerberg, I. K., Guerrero, J.-L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. E., & Xu, C.-Y. (2011). Calibration of hydrological models using flow-duration curves. *Hydrol. Earth Sys. Sci.*, 15(7), 2205-2227. <http://dx.doi.org/10.5194/hess-15-2205-2011>.
- Willmott, C. J. (1981). On the validation of models. *Phys. Geography*, 2(2), 184-194.
- Willmott, C. J. (1984). On the evaluation of model performance in physical geography. In *Spatial Statistics and Models* (pp. 443-460). Springer.
- Yuan, Y., Khare, Y., Wang, X., Parajuli, P. B., Kisekka, I., & Finsterle, S. (2015). Hydrologic and water quality models: Sensitivity. *Trans. ASABE*, 58(6), 1721-1744. <http://dx.doi.org/10.13031/trans.58.10611>.
- Zeckoski, R. W., Smolen, M. D., Moriasi, D. N., Frankenberger, J. R., & Feyereisen, G. W. (2015). Hydrologic and water quality terminology as applied to modeling. *Trans. ASABE*, 58(6), 1619-1635. <http://dx.doi.org/10.13031/trans.58.10713>.
- Zheng, C., Hill, M. C., Cao, G., & Ma, R. (2012). MT3DMS: Model use, calibration, and validation. *Trans. ASABE*, 55(4), 1549-1559. <http://dx.doi.org/10.13031/2013.42263>.



Appendix

ADAPT ♦ ♦ ♦ Agricultural Drainage and Pesticide Transport

AGWA ♦ ♦ ♦ ArcGIS-based Automated Geospatial Watershed Assessment

APEX ♦ ♦ ♦ Agricultural Policy/Environmental eXtender

BASINS ♦ ♦ ♦ Better Assessment Science Integrating Point and Nonpoint Sources

COUPMODEL ♦ ♦ ♦ Coupled Heat and Mass Transfer model

CREAMS ♦ ♦ ♦ Chemicals, Runoff, and Erosion from Agricultural Management Systems

DAISY ♦ ♦ ♦ Danish Simulation Model

EPIC ♦ ♦ ♦ Erosion Productivity Impact Calculator

GLEAMS ♦ ♦ ♦ Groundwater Loading Effects of Agricultural Management Systems

HSPF ♦ ♦ ♦ Hydrological Simulation Program - Fortran

H/WQ Hydrologic and water quality (models)

HYDRUS-

KINEROS KINematic runoff and EROSion

MIKE SHE MIKE System Hydrologique European (SHE)

MT3DMS Modular 3-Dimensional Multispecies Transport Model

RZWQM Root Zone Water Quality Model

SHAW Simultaneous Heat and Water

STANMOD Studio of ANalytical MODEls

SWAT Soil and Water Assessment Tool

SWIM Soil Water Infiltration and Movement

TOUGH Transport of Unsaturated Groundwater and Heat

VS2DI-

WAM Watershed Assessment Model

WARMF Watershed Analysis Risk Management Framework

WEPP Water Erosion Prediction Project

d Index of agreement

D_v Deviation volume

HUC Hydrologic unit code

MAE Mean absolute error

ME Mean error

MSE Mean square error

NSE Nash-Sutcliffe efficiency

PBIAS Percent bias

PE Prediction error

PPS Point to plot scale

PVE Percent volume error

r Pearson's correlation coefficient

R^2 Coefficient of determination

RB Relative bias

RE Relative error

RMSD Root mean square deviation

RMSE Root mean square error

RSR RMSE-observations standard deviation ratio

RVE Relative volume error

SD Standard deviation

ASABE Site Map

- [Publications](#)
- [Membership](#)
- [Meetings & Events](#)
- [Standards](#)
- [Career Resources](#)
- [Media](#)
- [Contact Us](#)
- [About Us](#)
- [News & Public Affairs](#)
- [eForums](#)
- [Foundation](#)
- [Awards & Landmarks](#)
- [Technical Library](#)

American Society of Agricultural and Biological Engineers

2950 Niles Road, St. Joseph, MI 49085

Phone: (269) 429-0300 Fax: (269) 429-3852 hq@asabe.org

Copyright 2016 American Society of Agricultural and Biological Engineers